The University Of Sheffield.

Access to Electronic Thesis

Author:	George Burghel
Thesis title:	Inherited and Somatic Genetic Factors in Colorectal Cancer Development
Qualification:	PhD

This electronic thesis is protected by the Copyright, Designs and Patents Act 1988. No reproduction is permitted without consent of the author. It is also protected by the Creative Commons Licence allowing Attributions-Non-commercial-No derivatives.

This thesis was embargoed until October 2013.

If this electronic thesis has been edited by the author it will be indicated as such on the title page and in the text.

INHERITED AND SOMATIC GENETIC FACTORS IN COLORECTAL CANCER DEVELOPMENT

BY
GEORGE J. BURGHEL

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

The
University
Of
Sheffield.

Institute for Cancer Studies/ Department of Oncology

School of Medicine

September 2011

ACKNOWLEDGMENTS

First and foremost, I would like to give my deepest appreciation to my supervisor, Dr Angela Cox, for her sound advice, patience and help throughout the project. I would also like thank Dr Patrick Eyers for his valuable comments on the thesis.

I am very thankful to everybody in our research group for their help throughout the project. I would especially like to thank Dr Wei-Yu Lin and Dan Connely for their help with data analysis, Dr Ian Brock, Dr Marina Parry and Dr Sushila Rigas for their support in the lab and Helen Cramp for collecting patient data. I am also grateful to Dr Dave Hammond for his time and effort spent in discussing aCGH data together. I would like to extend my thanks to everybody in the Institute for Cancer Studies, especially Dr Pedro Zuazua-Villar, Anil Ganesh and Sheila Blizard for their help in the lab. I would also like to thank my friends and family in England, especially Peter Jones, Jonathan Norgate, Ammar Adnan, Ewa Dudziec and the Parry's. Special thanks to my closest friends in Jordan, Amal Afifi, Muna Said, Fadi Sawaqedi, Noor Gammoh and Mohammed Edelbi. I am also very thankful to the University of Sheffield for their financial support and for offering me such a precious experience.

To my Fiancée, Heather, and my beloved family in Jordan, my dad "Joseph", my brothers Abboud and Rami and my sister in law Hanan, I truly cannot thank you enough for all your love and support throughout. I could have never done it without you.

Finally, I would like to dedicate this work to my loving mother Diana "Em Abboud", who is bravely fighting Glioblastoma Multiforme with great spirit and amazing positivity that was an inspiration for everybody around her.

LIST OF ABBREVIATIONS

Amino acids

Amino acid	Three letter code	Single letter code
Alanine	Ala	A
Arginine	Arg	R
Asparagine	Asn	N
Aspartic acid	Asp	D
Cysteine	Cys	C
Glutamic acid	Glu	E
Glutamine	Gln	Q
Glycine	Gly	G
Histidine	His	H
Isoleucine	Ile	I
Leucine	Leu	L
Lysine	Lys	K
Methionine	Met	M
Phenylalanine	Phe	F
Proline	Pro	P
Serine	Ser	S
Threonine	Thr	T
Tryptophan	Trp	W
Tyrosine	Tyr	Y
Valine	Val	V

A	Adenine
aa	Amino Acid
aCGH	Array Comparative Genome Hybridisation
ADM	Aberration Detection Method
APC	Adenomatous Polyposis Coli
APS	Ammonium persulfate
ATTC	American Type Tissue Collection
BAC	Bacterial Artificial Chromosome
BLAST	Basic Local Alignment Search Tool
BRAF	v-raf Murine Sarcoma Viral Oncogene Homolog B1 gene
BSA	Bovine Serum Albumin
C	Cytosine
CAE	Cancer Associated Event
Caspase	Cysteinyl Aspartate-Specific Protease
CASP8	Caspase 8 gene
CGH	Comparative Genome Hybridisation
Chr	Chromosome

CIMP	CpG Island Methylator Phenotype
CIN	Chromosomal Instability
CNA	Copy Number Aberration
CNV	Copy Number Variant
COCA	Context Corrected Common Aberration
dbSNP	The Single Nucleotide Polymorphism Database
dd	Double Distilled
Df	Degree of freedom
Del	Deletion
cDNA	Complimentary DNA
DISC	Death-Inducing Signalling Complex
DMEM	Dulbecco's Modified Eagle Medium
DNA	Deoxyribonucleic Acid
dNTPs	Deoxynucleotide triphosphates
DSB	Double-Strand Break
EDTA	Ethylenediamine tetra acetic acid
EGFR	Epidermal Growth Factor Receptor
Et.Br.	Ethidium Bromide
Exo I	Exonuclease I
FAP	Familial Adenomatous Polyposis
FCS	Foetal Calf Serum
FE	Feature Extraction
FISH	Fluorescence *In Situ* Hybridization
FMCR	Focal Minimal Common Region
G	Guanine
GBM	Glioblastoma Multiforme
GP	General Practice
GWAS	Genome Wide Association Study
HCl	Hydrochloric Acid
Hg	Human genome
HGVS	Human Genome Variation Society
HNPCC	Hereditary Nonpolyposis Colorectal Cancer
IARC	International Agency for Research on Cancer
Ins	Insertion
IU/dl	International Unit per Decilitre

Kb	Kilobase
KRAS	v-Ki-ras2 Kirsten Rat Sarcoma Viral Oncogene Homolog gene
Lab	Laboratory
LOH	Loss of Heterozygousity
MCR	Mutation Cluster Region
mg	Milligram
$MgCl_2$	Magnesium Chloride
min	Minute
ml	Millilitre
µl	Microlitre
µM	Micro molar
MMR	Mismatch Repair
mRNA	Messenger RNA
MSI	Microsatellite Instability
MS-MLPA	Methylation Specific Multiplex Ligation Probe Amplification
MSP	Methylation Specific PCR
MSS	Microsatellite Stable
NaAc	Sodium Acetate
N/A	Not Applicable
NCBI	National Center for Biotechnology Information
NCI	National Cancer Institute
ng	Nanogram
NGRL	National Genetics Reference Laboratories
NTC	No Template Control
PAGE	Polyacrylamide Gel Electrophoresis
PBS	Phosphate Buffer Saline
PCR	Polymerase Chain Reaction
PIK3CA	Phosphoinositide-3-Kinase Catalytic Alpha Polypeptide gene
QF-PCR	Quantitative Fluorescent PCR
RNA	Ribonucleic Acid
RPM	Round Per Minute
RT-PCR	Real Time PCR
SAP	Shrimp Alkaline Phosphatase
SDE	Size Determining Event
SDS	Sodium Dodecyl Sulfate

sec	Second
SNP	Single Nucleotide Polymorphism
STR	Short Tandem Repeat
T	Thymine
Taq	*Thermus aquaticus*
TBE	Tris-borate EDTA
TE	Tris-EDTA
TEMED	Tetramethylethylenediamine
T_M	Melting Temperature
TP53	Tumour Protein p53 gene
TSG	Tumour Suppressor Gene
u	Unit
UCR	Utah Cancer Registry
UK	United Kingdom
UPDB	Utah Population Database
USA	United States of America
3’ UTR	3-prime Untranslated Region
UV	Ultra-violet
v	Volume
w	Weight
WGA	Whole Genome Amplification

ABSTRACT

Colorectal cancer (CRC) is the 3rd most common cancer and the 4th highest cause of cancer deaths in the world. Genetic factors play a major role in its predisposition, initiation and development. Inherited variants in the *CASP8* gene, a key regulator of apoptosis, have a potential yet controversial association with CRC risk. Sporadic CRC develop through different molecular pathways of genomic instabilities and mutations in key cancer driver genes. Classification of sporadic CRC into these molecular pathways has potential implications for diagnosis and treatment and it is an integral part of CRC studies, however, current published research suffers from lack of standardisation. Chromosomal Instability (CIN) drives CRC by affecting cancer driver genes, many of which are still to be identified.

This project aimed to: (a) further investigate the role of *CASP8* inherited variants in CRC risk, (b) to molecularly classify sporadic CRC tumour DNA samples using standard techniques and definitions, and (c) to identify novel CRC driver genes affected by CIN. A *CASP8* promoter in/del variant was genotyped in 1193 CRC cases and 1388 matching controls. The coding region of the *CASP8* gene was sequenced in 94 CRC cases to identify potential novel variants and a copy number variant was also investigated. A cohort of 53 paired CRC tumour and normal DNA samples were molecularly classified using standard techniques and definitions. Common aberration analysis was performed on high resolution array comparative genome hybridisation data from 45 chromosomally unstable CRC cases to identify focal minimal common regions (FMCR).

CASP8 inherited variants did not significantly affect CRC risk in the investigated cohort. CRC molecular classification confirmed the heterogeneity of sporadic CRC and a novel molecular subtype was proposed. FMCR were shown to target cancer related genes and novel CRC driver genes were proposed. Finally, preliminary studies supported the tumour suppressor role of *NFKBIA*, one of the novel candidate driver genes affected by a deletion FMCR.

Publications (in preparation)

Burghel GJ, Lin W-Y, Connely D, Brock IW, Hammond D, Cross SS, Bury J, Stephenson Y, Cox A. Identification of candidate driver genes in common focal chromosomal aberrations of microsatellite stable colorectal cancer.

Curtin K, Knight S, Abo R, Cai Z, Cannon-Albright LA, Northwood E, Bishop DT, Lin W-Y, **Burghel GJ**, Camp NJ, Cox A. CASP8 variants are associated with female early onset colon cancer.

CONTENTS

LIST OF TABLES

LIST OF FIGURES

1. INTRODUCTION

1.1 Cancer

Cancer is a complex group of distinct syndromes characterised by uncontrolled growth and spread of malignant cells. Tumours may affect different organs of the body and each tumour can be divided into several types and subtypes. More than 100 distinct types of cancer are currently known (Hanahan and Weinberg 2000, Stratton *et al.* 2009). Worldwide, cancer is responsible for around 1 in 8 deaths, with an estimated total of 7.6 million deaths in the year 2008 (Stratton *et al.* 2009, Jemal *et al.* 2011). This makes cancer the 3rd leading cause of death after cardiovascular and infectious diseases (WHO 2008). In general, the overall lifetime risk of developing cancer is around 50% for men and 33% for women (based on data from the years 2001-2003) (Hayat *et al.* 2007).

1.2 The genetics of cancer

Genetic factors play a primary role in the predisposition, initiation and development of cancer (de la Chapelle 2004, Vogelstein and Kinzler 2004, Stratton *et al.* 2009). The familial clustering of cancer cases was firstly described by the French physician Paul Broca, who observed that his wife's family had multiple cancer cases (mainly of the breast) over 4 generations (Steel *et al.* 1991). Early suggestions for the involvement of genomic abnormalities in the development of cancer were made by David von Hansemann in the year 1890 and Theodor Boveri in the year 1902 when they noticed an "abnormal chromosome constitution" in malignant tumours and cancer cells (Marte 2006, Harris 2008). After the discovery of DNA as the hereditary material in chromosomes, these early suggestions of a genetic involvement in tumourigenesis were supported by observations that DNA damaging agents can result in the development of cancer (Loeb and Harris 2008, Stratton *et al.* 2009). Currently, germline mutations are known to be responsible for hereditary cancer predisposition and germline variants are known to affect both hereditary and sporadic cancer risk (Groden *et al.* 1991, Miki *et al.* 1994, Amundadottir *et al.* 2006, Easton *et al.* 2007, Tomlinson *et al.* 2008). Moreover, acquired genetic and genomic changes are known to drive cancer development (Vogelstein *et al.* 1988, Sjoblom *et al.* 2006, Wood *et al.* 2007, Leary *et al.* 2008, Beroukhim *et al.* 2010)

1.2.1 Germline genetic factors and cancer predisposition

Cancer predisposing germline genetic factors can be divided into 3 main types; rare, highly penetrant mutations, rare, moderate penetrance mutations and common, lower penetrance variants (Balmain *et al.* 2003, de la Chapelle 2004). Highly penetrant germline mutations have been mostly identified through linkage studies in affected families. They include mutations in tumour suppressor genes (TSG) such as adenomatosis polyposis coli (*APC*), responsible for familial adenomatous polyposis (FAP), and the breast cancer associated genes (*BRCA1* and *BRCA2*) responsible for familial breast cancer cases (Groden *et al.* 1991, Miki *et al.* 1994, Wooster *et al.* 1995). These highly penetrant genetic factors are rare, accounting for ~20% of familial cancer cases and ~5% of cancer cases in general (Balmain *et al.* 2003, de la Chapelle 2004). Moderate penetrance mutations also occur in TSG and were mainly identified by sequencing candidate genes in familial cancer cases or association studies (Meijers-Heijboer *et al.* 2003, Cybulski *et al.* 2004). *CHEK2* mutations, for example, confer moderate risk for several cancer types (Cybulski *et al.* 2004). However, they are also rare and account for <5% of cancer cases (Hemminki *et al.* 2009a, Hemminki *et al.* 2009b).

Much of the unknown inheritance can probably be explained by interactions between the more common but less penetrant genetic (and environmental) factors, based on a common variant common disease model (Balmain *et al.* 2003, Pasche and Yi 2010). These common, low penetrance genetic factors cannot be identified by linkage studies, but instead require large case control association studies (de la Chapelle 2004). Two different approaches of association studies have helped in the identification of several common cancer susceptibility genes with low penetrance; candidate gene association studies and genome wide association studies (GWAS) (Pasche and Yi 2010, Chung and Chanock 2011). Candidate genes are usually selected based on their involvement in cancer related pathways such as cell proliferation and apoptosis (Pasche and Yi 2010). On the other hand, GWAS search for common susceptibility factors across the whole genome. GWAS are based on high throughput technologies, which genotype hundreds of thousands of common single nucleotide polymorphisms (SNPs) in large numbers of cancer cases and controls (Chung and Chanock 2011). Recently, copy number variants (CNVs) became of use in association studies as an alternative for SNPs (Beckmann *et al.* 2007, Venkatachalam *et al.* 2010).

1.2.2 Genetic instability, somatic changes and cancer development

Throughout the development of cancer, cancer genomes acquire various somatic genetic changes. These changes can occur on several levels, ranging from single nucleotide mutations to the deletion or amplification of entire chromosomes (Balmain *et al.* 2003). Some of these somatic changes play an important role in cancer development, usually by affecting key cancer driver genes. On the other hand, some of the changes do not provide any selective advantage and are termed passenger mutations (Greenman *et al.* 2007). Despite the large number of cancer types and subtypes (>100), cancers are known to share a common pathogenesis in which they acquire 6-8 capabilities known as the hallmarks of cancer. These hallmarks are shared, more or less, by all types of cancer and they enable the progressive transformation of cells from normal to malignant (Hanahan and Weinberg 2011). The hallmarks include the ability of the cancer cells to: induce and sustain proliferative signalling, avoid growth suppression, resist cell death and apoptosis, replicate continually, induce angiogenesis, activate invasion and metastasis, avoid immune destruction and deregulate cellular bioenergetics (Hanahan and Weinberg 2000, Hanahan and Weinberg 2011).

Cancer hallmarks are mainly acquired through multiple steps of somatic genetic and epigenetic alterations and mutations affecting cancer driver genes (Loeb 1991, Hanahan and Weinberg 2011). Normally, the human genome is stable and is maintained by DNA monitoring and repair pathways which ensure that mutations are very rare events. If the mutations responsible for tumourigenesis were to occur at these normal rates, cancer would be a very rare disease (Loeb 1991). However, cancers occur at substantial frequency worldwide. For example, in the year 2008, ~12.7 million new cancer cases were reported (Jemal *et al.* 2011). Therefore, in order to acquire enough mutations and occur at such high incidence, cancers must be genetically unstable. This is why genomic instability is considered one of the main driving forces behind cancer development (Loeb 1991, Hanahan and Weinberg 2000, Hanahan and Weinberg 2011).

1.3 Colorectal cancer

Colorectal cancer (CRC) is the third most common cancer in males (Figure 1.1) and the second in females (Figure 1.2) (Jemal *et al.* 2011). More than 1 million new CRC cases are diagnosed annually and ~600,000 related deaths were estimated worldwide in the year 2008, making CRC the 3rd highest cause of cancer related death in both genders (Figure 1.3) (Parkin *et al.* 2005, Jemal *et al.* 2011). CRC is multifactorial and several environmental and genetic factors contribute to its aetiology (Jong *et al.* 2002). Environmental factors include diet, smoking and physical activity. In general, obesity, smoking and diets rich in red meat are associated with higher risk while regular physical exercise and diets rich in folate and calcium are associated with lower risk (Jong *et al.* 2002).

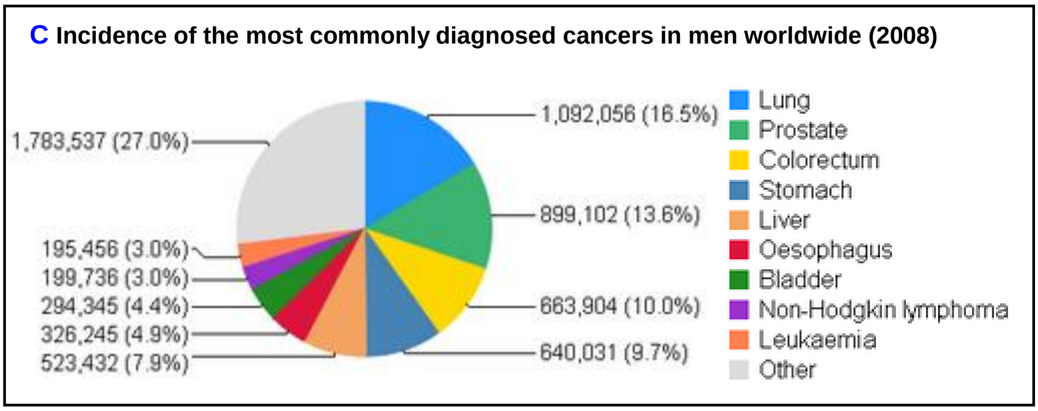


Figure 1.1 Worldwide cancer incidence in males

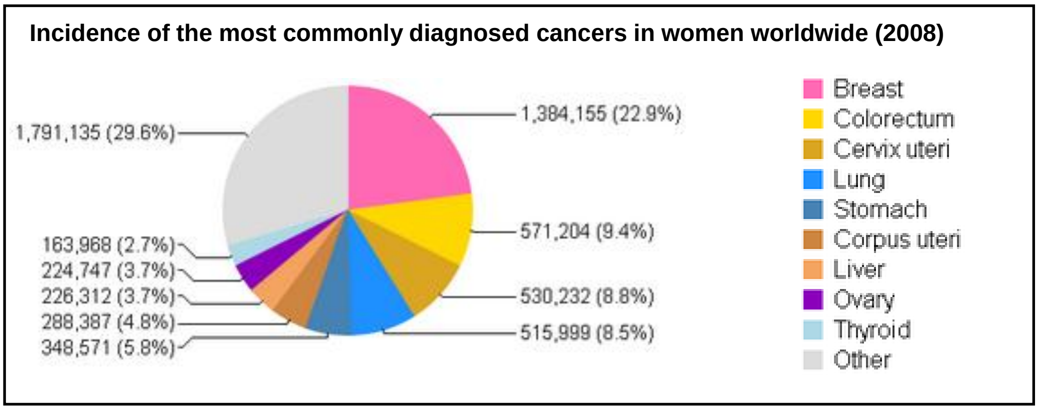


Figure 1.2 Worldwide cancer incidence in females

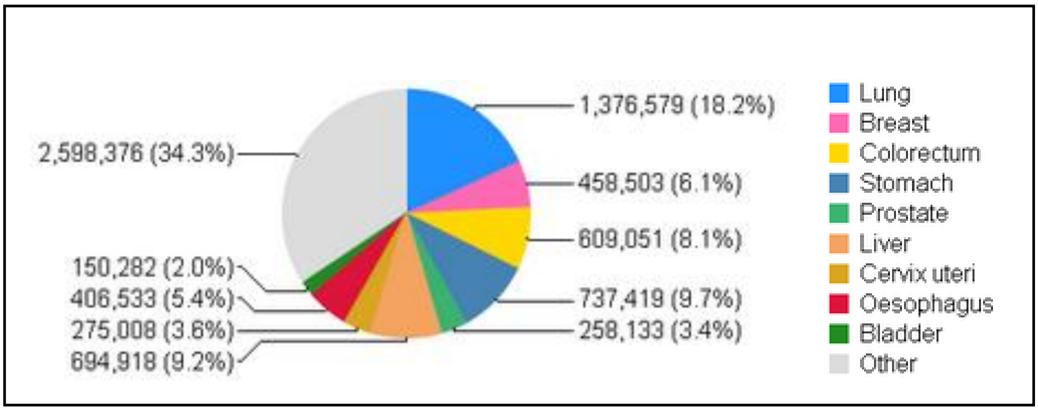


Figure 1.3 Worldwide cancer mortality in males and females

Figures 1.1-1.3 were generated based on data from the GLOBOCAN 2008 project available from the International Agency for Research on Cancer (IARC) website http://globocan.iarc.fr/ (accessed July, 2011).

1.3.1 Colorectal cancer: pathology and staging

CRC progresses slowly over years and tumour development occurs through several distinct histo-pathological stages. These stages range from single crypt lesions through small adenomatous polyps (benign tumours) to malignant carcinomas with metastatic potential (Deschoolmeester *et al.* 2010, Migliore *et al.* 2011). CRC stages are mainly defined by the tumour, node, metastasis (TNM) staging system of the American Joint Committee on Cancer (AJCC) and the International Union Against Cancer (UICC), or the Dukes staging system (Compton and Greene 2004, Migliore *et al.* 2011). TNM staging is based on the local extent and the depth of invasion of the primary tumour (T), the level of lymph node involvement (N) and the presence of distant metastatic cancer (M) (Compton and Greene 2004). The Dukes staging system, which has undergone several modifications since it was originally described by the British pathologist Cuthbert Dukes in 1932, includes the parameters observed in TNM staging, but also describes the level of cancer cell differentiation (Jass and Morson 1987, Migliore *et al.* 2011). In general, both staging systems divide CRC into 4 main stages that reflect the severity and development of CRC (Table 1.1). Although this histo-pathological staging of CRC is useful in stratifying patients into distinct groups to predict prognosis and survival, the system is limited in predicting treatment response or defining the underlying mechanisms of tumourigenesis (Westra *et al.* 2005, Ogino and Goel 2008, Walther *et al.* 2009).

Table 1.1 Summary of TNM and modified Duke's staging systems

TNM Stages		**Modified Duke's**	**Survival (%)***
Stage 0	Tis, N0, M0	-	-
Stage I (no nodal involvement, no metastases)	T1 (submucosa invaded)	A	90-100
	T2 (muscularis propria invaded)	B1	
Stage II (no nodal involvement, no metastases)	IIA (T3, subserosa invaded)	B2	75-85
	IIB (T4, visceral peritoneum invaded)	B3	
Stage III (no metastases)	IIIA (T1-T2, N1 (1-3 nodes involved))	C1	30-40
	IIIB (T2-T4, N1 (1-3 nodes involved))	C2/C3	
	IIIC (T1-T4, N2 (≥4nodes involved))	C1-C3	
Stage IV (distant metastases)	T1-T4, N1-N4, M1	D	<5

* 5 year survival estimates

Table 1.1 was adapted from (Kerr D J 2001, Horton and Tepper 2005).

1.3.2 Colorectal cancer: genetics

CRC is divided into sporadic and hereditary (familial) cases with genetic factors playing a significant role in predisposition to the disease (de la Chapelle 2004). Around 75% of CRC cases are sporadic and the remaining 25% of cases show hereditary or familial CRC characteristics (Migliore *et al.* 2011). Large cancer heritability studies based on 1st degree family members have shown that amongst common cancers, CRC has one of the highest genetic contributions towards its risk (Goldgar *et al.* 1994, Dong and Hemminki 2001). Moreover, a large multi-cohort twin study has estimated that the genetic contribution in CRC is ~35%, making it the 2nd highest (behind prostate cancer) in terms of heritability of the common cancers (Lichtenstein *et al.* 2000, de la Chapelle 2004). Nevertheless, known and highly penetrant inherited genetic factors that predispose to familial CRC are rare and account for only 5-6% of cases (Migliore *et al.* 2011). The two main inherited and well-characterised forms of CRC are Lynch syndrome and familial adenomatous polyposis (FAP) (Lynch and de la Chapelle 2003, de la Chapelle 2004).

1.3.2.1 Lynch syndrome

Lynch syndrome or hereditary non-polyposis colorectal cancer (HNPCC) was originally described in 1913 by the pathologist Aldred Warthin, who investigated a family with a high incidence of cancer (Lynch and de la Chapelle 1999). The study of this family, known as family “G” took place over several decades, with the disease being described as a cancer family syndrome in the mid 1960’s (Lynch *et al.* 1966). In 1970, family “G” (with more than 650 blood relatives and 95 cancer cases) was restudied in detail by Henry Lynch and Anne Krush who described an autosomal dominant hereditary cancer syndrome (Lynch and Krush 1971).

Lynch syndrome is mainly caused by germline mutations in one of several DNA mismatch repair (MMR) genes such as *MLH1*, *MSH2*, *MSH6* and *PMS2* (de la

Chapelle 2004, Migliore *et al.* 2011). Loss of MMR gene function results in microsatellite instability (MSI) (Section 1.3.3.2) which occurs in ~90% of Lynch syndrome colorectal tumours (Nussbaum *et al.* 2001). The syndrome is inherited in an autosomal dominant pattern and accounts for ~3% of the total CRC cases (Nussbaum *et al.* 2001, Migliore *et al.* 2011). People carrying a germline mutation in one of the MMR genes will have ~80% risk of developing CRC (de la Chapelle 2004, Migliore *et al.* 2011). In addition to CRC, Lynch syndrome patients are at an increased risk of other forms of cancer, including those of the stomach, ovaries, brain endometrium and pancreas (Lynch and de la Chapelle 2003, de la Chapelle 2004, Migliore *et al.* 2011). In 60-70% of Lynch syndrome cases, the tumour is proximal to the splenic flexure (Lynch and de la Chapelle 1999, Nussbaum *et al.* 2001).

1.3.2.2 Familial adenomatous polyposis

Familial adenomatous polyposis (FAP) is characterised by the formation of hundreds, even thousands, of adenomas in the colorectum, usually starting early in life (occasionally in the preteen years) (Nussbaum *et al.* 2001, Lynch and de la Chapelle 2003). FAP is caused by mutations in a TSG, namely adenomatous polyposis coli (*APC*). FAP is an autosomal dominant syndrome, and accounts for ~1% of CRC cases (Migliore *et al.* 2011). The penetrance of FAP is almost 100% and CRC development is inevitable (at the age of 40 to 50 years) if the adenomas are left untreated (Lynch and de la Chapelle 2003, Migliore *et al.* 2011). Therefore, individuals who have a first-degree relative with FAP should be screened for *APC* mutations or have flexible sigmoidoscopy early in life (Lynch and de la Chapelle 2003). Any patients carrying germline *APC* mutations or who have colonic polyps require an annual endoscopy examination (Lynch and de la Chapelle 2003). It is worth mentioning that mutations in *APC* occur in up to 70% of sporadic colorectal tumours (Miyaki *et al.* 1994, Luchtenborg *et al.* 2004).

1.3.2.3 Low penetrance genes and colorectal cancer risk

As stated earlier, Lynch syndrome and FAP are caused by highly penetrant and well characterised mutations. However, these syndromes are rare, and account for <5% of all CRC cases and only ~20% of hereditary CRC cases (Jong *et al.* 2002, de la Chapelle 2004). The remaining majority of the hereditary cases, and a proportion of the sporadic CRC cases, are likely to be caused by more common but less penetrant genetic variants and polymorphisms (SNPs) (Broderick *et al.* 2007). These are generally difficult to identify as linkage analysis studies do not have enough power to detect low penetrance variants (de la Chapelle 2004). Nevertheless, genetic association studies using both the candidate gene and the genome wide approaches have helped to identify some of these less penetrant genetic variants (de la Chapelle 2004, Migliore *et al.* 2011).

Some of the genetic variants identified through candidate gene association studies occur within the genes of the metabolic enzymes Glutathione S-transferases (GSTs) methylenetetrahydrofolate reductase (*MTHFR*), and DNA repair genes such as *XRCC2* (Huang *et al.* 2007, Curtin *et al.* 2009b, Economopoulos and Sergentanis 2010). However, most of these associations are yet to be confirmed in large and independent association studies (Migliore *et al.* 2011). Several of the genetic variants identified through GWAS occur within genes and loci that include *SMAD7*, *CRAC1 (HMPS)*, 8q23.3, 8q24, 10p14, 1q41, 3q26.2, 12q13.13 and 20q13.33 (Broderick *et al.* 2007, Haiman *et al.* 2007, Tomlinson *et al.* 2007, Zanke *et al.* 2007, Jaeger *et al.* 2008, Tenesa *et al.* 2008, Tomlinson *et al.* 2008, Houlston *et al.* 2010). Recently, using fine mapping and a candidate gene approach based on GWAS results, genetic variants within *GREM1*, *BMP4* and *BMP2* were identified as CRC risk factors (Tomlinson *et al.* 2011). In general, all the verified common risk variants account for ~6% of the unknown CRC heritability (Lascorz *et al.* 2010). The identification of more of these lower penetrance genetic variants will play an important role in increasing the knowledge of the genetic pathways involved in the initiation, development and progression of CRC (de la Chapelle 2004). Moreover, this knowledge also has possible implications for the prevention, diagnosis and treatment of CRC (de la Chapelle 2004, Lerman and Shields 2004).

1.3.3 Genomic instability and development of colorectal cancer

As discussed earlier, genomic instability is a primary driving force in cancer development. Three main forms of genomic instability occur in CRC; chromosomal instability (CIN), microsatellite instability (MSI) and epigenetic instability, known as CpG island methylator phenotype (CIMP) (Thibodeau *et al.* 1993, Toyota *et al.* 1999, Pino and Chung 2010, Migliore *et al.* 2011). Studying the global genomic status of CIN, MSI and CIMP can play an important role in determining the biological and clinico-pathological characteristics of CRC (Issa 2008, Ogino and Goel 2008).

1.3.3.1 Chromosomal instability

1.3.3.1.1 The basis of chromosomal instability

CIN is the most common type of genomic instability in CRC and accounts for 65-85% of cases (Derks *et al.* 2008, Issa 2008, Pino and Chung 2010, Migliore *et al.* 2011). Tumours that develop through the CIN pathway are characterised by frequent numerical and structural gains and losses of chromosomal fractions or whole chromosomes at a significantly increased rate in comparison to normal cells (Rajagopalan *et al.* 2003). CIN is thought to play an important role in tumourigenesis through the amplification of oncogenes such as *MYC* and the deletion of TSG such as *SMAD4* and *TP53* (Kozma *et al.* 1994, Ozakyol *et al.* 2006, Tanaka *et al.* 2006, Ogino and Goel 2008, Migliore *et al.* 2011). The molecular basis of CIN is not well understood. However, it can result from defects in the pathways which regulate and maintain chromosome segregations including the DNA damage response, the spindle assembly checkpoints, DNA double-strand break (DSB) repair and telomere regulation (Zhivotovsky and Kroemer 2004, Pino and Chung 2010). Mutations in genes that play a role in regulating these pathways, such as *APC*, *MRE11* and *TP53*, are implicated in CRC development and are associated with CIN (Wang *et al.* 2004, Pino and Chung 2010). Another important cause of CIN, especially with numerical chromosomal defects (rather than structural), is abnormal centrosome function. *AURKA*, an important regulator of centrosome function, is overexpressed in several cancers and was recently shown to be associated with CIN in CRC (Baba *et al.* 2009).

1.3.3.1.2 Characterisation of chromosomal instability

Several techniques and approaches are available to characterise and define CIN. They include conventional methods such as karyotyping, loss of heterozygousity (LOH) analysis using microsatellite markers, DNA ploidy analysis using flow cytometry, fluorescent *in situ* hybridisation (FISH) and conventional comparative genome hybridisation (CGH) and high throughput methods such as bacterial artificial chromosome (BAC) CGH and oligonucleotide array CGH and SNP arrays (Hermsen *et al.* 2002, Chang *et al.* 2006, Cheng *et al.* 2008, Derks *et al.* 2008, Baba *et al.* 2009, Pino and Chung 2010, Poulogiannis *et al.* 2010a). These diverse approaches and techniques provide different criteria to define CIN; making it very difficult to compare results between studies or introduce standard definitions for CIN (Pino and Chung 2010). Nevertheless, LOH at several microsatellite markers, such as those on 18q, 8p and 5q, and the so-called cancer associated events (CAE) (losses in 8p21-pter, 15q11–q21, 17p12–13 and 18q12–21 and gains in 8q23-qter, 13q14–31 and 20q13) were proposed to represent CIN. However, these markers are not in standard use and are prone to underestimating CIN and producing false-negative results (Hermsen *et al.* 2002, Ogino and Goel 2008). High resolution and genome-wide array based techniques are currently the methods of choice; however, a standard definition of CIN is still lacking (Brosens *et al.* 2010, Pino and Chung 2010, Dyrso *et al.* 2011). More recently, next generation sequencing platforms are providing even higher resolution and accuracy in the detection of chromosomal aberrations in cancer

(Wong *et al.* 2011). However, their use is currently limited due to the higher cost involved.

1.3.3.2 Microsatellite instability

1.3.3.2.1 Microsatellite instability and CRC

In 1993, microsatellite instability (MSI), also known as the mutator phenotype, was identified in CRC as a molecular subtype with distinct genotypic and phenotypic characteristics (Ionov *et al.* 1993, Thibodeau *et al.* 1993). MSI is characterised by altered lengths of microsatellites in tumour DNA compared to normal DNA (Thibodeau *et al.* 1993). When microsatellites are located in coding regions of genes, MSI can result in frameshift mutations and altered protein function (Vilar and Gruber 2010). Some of the important cancer genes affected by MSI, and also associated with CRC include the pro-apoptotic genes *BAX*, *CASP5* and *BCL10*, the transforming growth factor beta receptor type II (*TGFBR2*), the insulin-like growth factor II receptor (*IGF2R*) the transcription factor *E2F*, the cell cycle control gene *CHEK1* and the DNA repair genes *MSH3* and *MSH6* (Yin *et al.* 1997, Bertoni *et al.* 1999, Trojan *et al.* 2004, Iacopetta *et al.* 2010, Yashiro *et al.* 2010).

1.3.3.2.2 Microsatellite instability: molecular basis and testing

MSI is associated with defective DNA mismatch repair machinery resulting from mutations or promoter hypermethylation of MMR genes such as *MLH1* and MSH2 (Zhivotovsky and Kroemer 2004, Ogino and Goel 2008). MSI is the main characteristic of Lynch syndrome, however, MSI only occurs in ~15% of sporadic colorectal tumours (Nussbaum *et al.* 2001). The global genomic status of MSI is assessed through the analysis of specific microsatellite markers. In 1997, the National Cancer Institute (NCI) set the guidelines for MSI testing and developed a panel of 5 microsatellite markers; 3 dinucleotide markers (D2S123, D5S346 and D17S250) and 2 mononucleotide markers (BAT25 and BAT26). These markers, called the NCI consensus panel or the Bethesda panel, were considered to be the reference panel for MSI testing (Boland *et al.* 1998). Based on this panel, tumours with instability in 2 or more markers were classified as MSI-high (MSI-H), while those with one unstable marker were defined as MSI-low (MSI-L) and tumours without any instability were considered microsatellite stable (MSS) (Boland *et al.* 1998, Bacher *et al.* 2004). However, the NCI consensus panel suffered from several limitations arising from the dinucleotide markers, which were shown to be less sensitive and specific for MSI testing when compared to mononucleotide markers (Suraweera *et al.* 2002).

In 2002, the NCI released the revised guidelines for MSI, and mononucleotide markers were recommended as the most sensitive markers for this type of testing (Umar *et al.* 2004). In 2004, a set of 5 mononucleotides markers (BAT25, BAT26, NR21, NR24 and MONO27) were developed for MSI testing according to the revised NCI criteria (Bacher *et al.* 2004). The latter panel was shown to be more accurate in determining the MSI-H status in comparison to the old panel (Murphy *et al.* 2006).

1.3.3.3 The CpG island methylator phenotype

Transcriptional silencing of TSG by DNA methylation at promoter CpG islands is very common in cancer cells (Ogino *et al.* 2006a). Several TSG, such as *CDKN2A* and *MLH1* were shown to be transcriptionally inactivated by DNA methylation in CRC (Ahuja *et al.* 1997, Kambara *et al.* 2004). In 1999, a CpG island methylator phenotype (CIMP) was proposed as a distinct form of epigenetic instability in a subset of CRC tumours (Toyota *et al.* 1999). CRC cases with CIMP were characterised by having a higher frequency of methylated TSG promoters in tumour DNA compared to DNA from normal colon tissues (Toyota *et al.* 1999, Toyota *et al.* 2000). The existence of CIMP as a unique form of epigenetic instability in CRC has now been established and confirmed (Ogino and Goel 2008).

1.3.3.3.1 CIMP testing

The global genomic status of CIMP is usually assessed by quantitatively analysing a set of CpG island methylation markers (Ogino and Goel 2008). Several marker panels have been proposed for the evaluation of CIMP in CRC. The classical panel that was initially used to investigate CIMP in CRC included the following genes: *CDKN2A*, *MINT1*, *MINT2*, *MINT31* and *MLH1* (Rashid *et al.* 2001, Hawkins *et al.* 2002, Frazier *et al.* 2003, Samowitz *et al.* 2005). However, other panels were later introduced and were proposed to be more robust than this classical group. One of these included the genes *CACNA1G*, *IGF2*, *NEUROG1*, *RUNX3* and *SOCS1* (Weisenberger *et al.* 2006, Cheng *et al.* 2008) and another included; *CACNA1G*, *CDKN2A*, *CRABP1*, *MLH1* and *NEUROG1* (Ogino *et al.* 2006a). In 2007, a large population study (based on 920 CRC samples) evaluated the performance of both panels and proposed the use of 4-8 genes (*RUNX3*, *CACNA1G*, *IGF2*, *MLH1*, *NEUROG1*, *CRABP1*, *SOCS1* and *CDKN2A*) as a highly sensitive and specific panel for the evaluation of CIMP in CRC (Ogino *et al.* 2007). However, the panel is still not in standard use (Kim *et al.* 2010, Jover *et al.* 2011).

Several techniques are also available to determine the methylation status of the CpG markers for the evaluation of CIMP. Some of the commonly used techniques include methylation specific PCR (MSP) (which are qualitative) and pyrosequencing and real time PCR based assays (that are quantitative) (Samowitz *et al.* 2005, Ogino *et al.* 2006a, Jover *et al.* 2011). In general, it is recommended that a minimum of 4-5 validated CIMP markers are employed, and avoiding the use of qualitative techniques (Ogino *et al.* 2006a, Ogino *et al.* 2007, Goel and Shin 2008). Like MSI, CIMP can be divided into several subgroups; intense methylation of several markers known as CIMP-high (CIMP-H), less extensive marker methylation (compared to CIMP-H) known as CIMP-low (CIMP-L) and absence of methylation or CIMP-negative (CIMP-N). Nevertheless, CIMP-L definition and its distinction from CIMP-H/N remain controversial (Ogino *et al.* 2006b, Goel and Shin 2008).

1.3.4 Genetic mutations and colorectal cancer

Germline mutations in genes such as *APC* and *MLH1* are known to predispose to hereditary CRC (Sections 1.3.2.1 and 1.3.2.2). On the other hand, somatic mutations in cancer related genes play a primary role (alongside genomic instabilities) in driving CRC development (Wood *et al.* 2007, Migliore *et al.* 2011). Some of the major oncogenes and TSG known to be mutated in sporadic CRC include *APC*, *TP53*, *KRAS*, *PIK3CA*, *BRAF* and *SMAD4* (Woodford-Richens *et al.* 2001, Davies *et al.* 2002, Brink *et al.* 2003, Iacopetta 2003, Luchtenborg *et al.* 2004, Ikenoue *et al.* 2005, Velho *et al.* 2005, Oliveira *et al.* 2007).

A comprehensive and systematic genome wide sequencing analyses of >18000 genes was recently performed in CRC and breast cancer (Sjoblom *et al.* 2006, Wood *et al.* 2007). Sequencing results from >100 CRC patients indicated the presence of ~80 somatic mutations in a typical colorectal tumour (Wood *et al.* 2007). However, statistical analysis predicted that <15 of these mutations were likely to be drivers of tumourigenesis (Wood *et al.* 2007). In total, 140 CRC driver genes were identified, and predicted to play an important role in CRC development and progression (Wood *et al.* 2007). Whilst including the well-known CRC driver genes *APC*, *TP53*, *KRAS*, *FBXW7*, *PIK3CA* and *SMAD4*, this exercise also identified a large number of genes that had never before been implicated in cancer. The results of these sequencing studies revealed the scale of somatic mutations in CRC and confirmed their central role in driving tumour development (Sjoblom *et al.* 2006, Wood *et al.* 2007).

1.3.5 Molecular classification & characterisation of colorectal cancer

As stated, CRC develops through a sequential and multistep process of accumulating genetic and epigenetic defects. Although these defects occur in a mostly random fashion, they tend to accumulate in patterns (Ogino and Goel 2008, Stratton *et al.* 2009). These organised patterns probably arise due to selection pressure created by advantages and/or disadvantages of carrying some of the molecular defects (Ogino and Goel 2008, Stratton *et al.* 2009), with advantageous defects providing the tumour cells with one or more of the cancer hallmarks explained earlier (Stratton *et al.* 2009). Molecular classification of CRC helps in understanding the patterns which underlie the molecular mechanisms of tumourigenesis. This in turn can help to identify molecular biomarkers with the potential to be used in the clinic for predicting treatment response and patients' prognosis or survival (Westra *et al.* 2005, Ogino and Goel 2008). In this way, molecular classification of CRC can complement histo-pathological classification (Section 1.3.1) and provide a more comprehensive picture that can be applied clinically.

For a long time it was considered that sporadic CRC develops through a uniform and linear pathway of molecular defects, starting with *APC* inactivating mutations and the subsequent acquisition of more defects such as *KRAS* activation and *TP53* inactivation (Fearon and Vogelstein 1990) (Figure 1.4). This homogenous view of sporadic CRC genetic development has directed both research and clinical management for several decades (Jass 2007, Issa 2008). However, more recently, integrated genetic and epigenetic studies on sporadic CRC cases have shown that it can no longer be considered as one uniform and homogenous disease (Shen *et al.* 2007, Cheng *et al.* 2008, Derks *et al.* 2008, Issa 2008). Instead, there are several subtypes of sporadic CRC (at least 3) which develop through different pathways of genetic and epigenetic instabilities (CIN, MSI and CIMP), and mutations in key driver genes such as; *APC*, *TP53*, *KRAS*, *BRAF* and *PIK3CA* (Figure 1.5) (Chang *et al.* 2006, Shen *et al.* 2007, Cheng *et al.* 2008, Derks *et al.* 2008, Issa 2008, Nosho *et al.* 2008). In addition to their distinct molecular characteristics, these molecular pathways/CRC subtypes are known to associate with different clinical and pathological features (Shen *et al.* 2007, Issa 2008) (Figure 1.5).

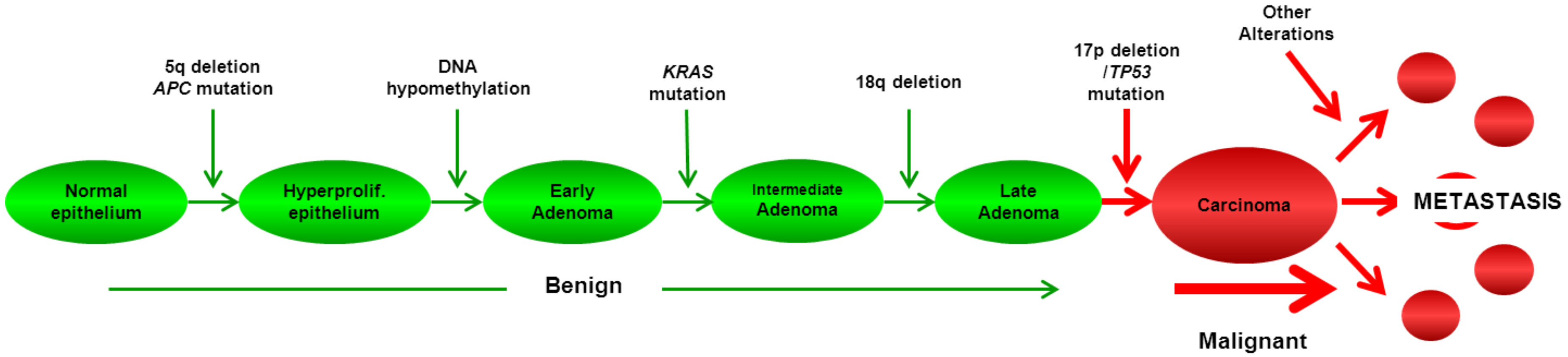


Figure 1.4 The classical linear model for sporadic CRC development

The classical linear model of genetic and epigenetic defects leading to CRC, the model is widely known as the “Vogelgram”.

Figure adapted from (Fearon and Vogelstein 1990).

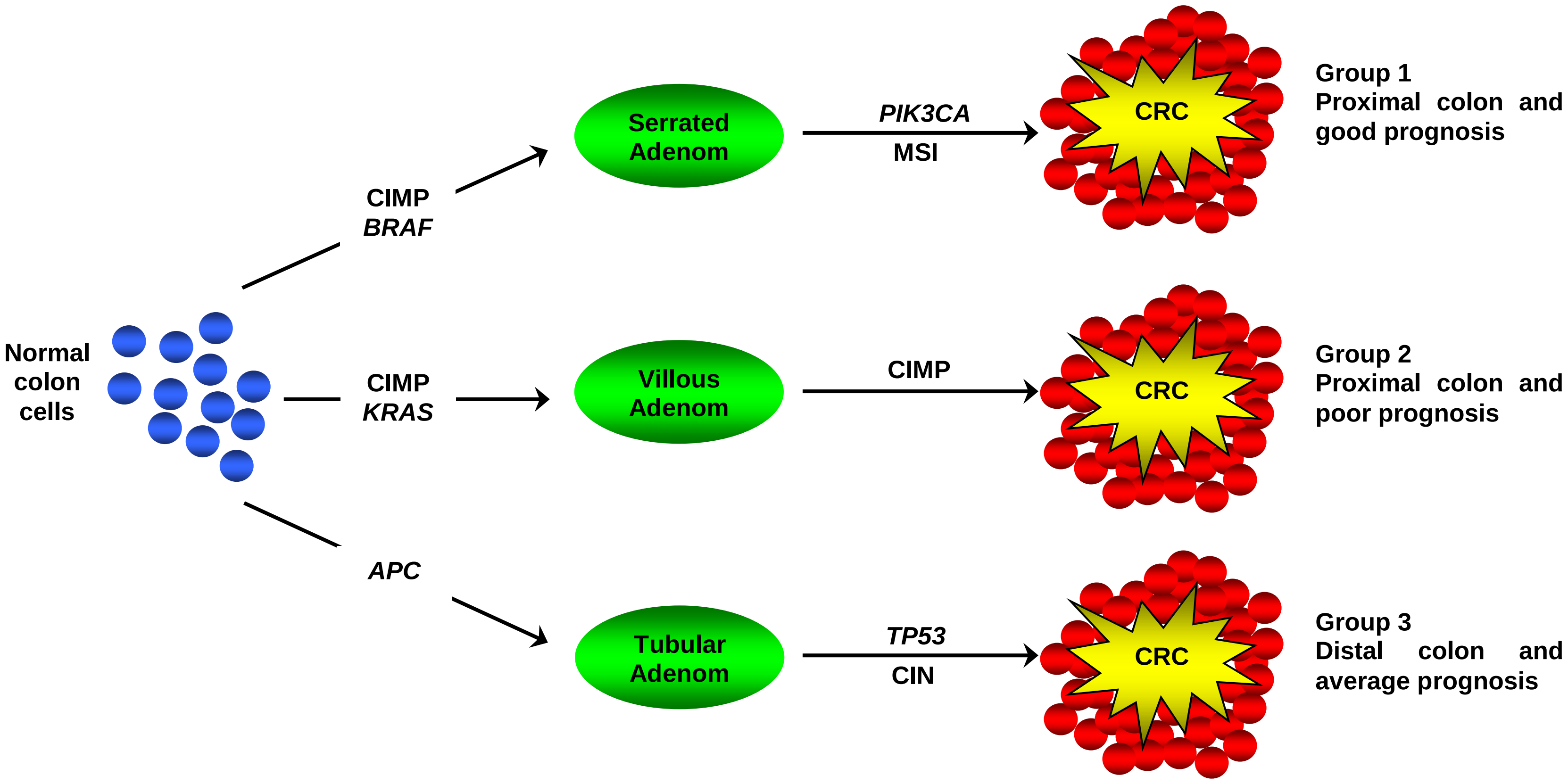


Figure 1.5 **Sporadic CRC molecular pathways**

Recent studies have shown that sporadic CRC cannot be considered as a single homogenous disease. There are at least 3 molecular pathways for CRC development. These pathways provide a better understanding of sporadic CRC development with potential applications in clinic (Adapted from (Issa 2008)).

1.3.5.1 An overview of CRC molecular pathways

Several comprehensive molecular classification systems for sporadic CRC have recently been proposed, based on genomic instabilities, genetic mutations and clinic-pathological features (Jass 2007, Issa 2008, Ogino and Goel 2008). The number of molecular subtypes ranges from 3-9 depending mainly on the genomic instabilities used and their definitions (Ogino and Goel 2008). Some of the proposed classification systems do not incorporate CIN (Ogino and Goel 2008). Moreover, some of the suggested subtypes are very rare, bearing subtle differences from the others which may not have any clinical relevance (Ogino and Goel 2008). The following sections (1.3.5.1.1-1.3.5.1.3) describe the 3 main subtypes based on a classification system proposed by Jean-Pierre Issa, which incorporates the 3 genomic instabilities (CIN, MSI and CIMP) (Issa 2008) (Figure 1.5).

1.3.5.1.1 The classical MSI-high pathway

The first molecular pathway of CRC development is known as the MSI pathway. This pathway accounts for 10-20% of sporadic CRC cases (Chang *et al.* 2006, Shen *et al.* 2007, Issa 2008). It is characterised by 2 genomic instabilities; high levels of MSI (MSI-H) and in most cases, intense methylation of a wide range of genes (CIMP-H) (Issa 2008). MSI in sporadic CRC is thought to occur as a result of CIMP affecting distinct MMR genes, mainly *MLH1* (Weisenberger *et al.* 2006). In addition to the genomic instabilities, this pathway is also characterised by mutations in key CRC driver genes. Mutations in *BRAF* (mainly V600E) occur at a frequency of 5-18% in sporadic CRC. However, they cluster within the MSI-H/CIMP-H pathway with a frequency of 40-50% (Oliveira *et al.* 2007, Shen *et al.* 2007). *PIK3CA* mutations were shown to play an important role in CRC development with an estimated frequency of 13-30% (Samuels *et al.* 2004, Ikenoue *et al.* 2005, Velho *et al.* 2005). Several studies have shown an association between the most common and oncogenic *PIK3CA* mutations and the MSI-H/CIMP-H sporadic CRC pathway (Abubaker *et al.* 2008, Nosho *et al.* 2008).

From a clinico-pathological perspective, sporadic colorectal tumours developing through the MSI-H/CIMP-H pathway are thought to arise from serrated or mucinous adenomas and mainly occur in the proximal colon (Jass 2007, Shen *et al.* 2007, Issa 2008). CRC patients with this pathway are more commonly females, and usually have good prognosis and survival (Samowitz *et al.* 2001a, Shen *et al.* 2007, Ogino and Goel 2008, Ogino *et al.* 2009).

1.3.5.1.2 The CIN positive/MSS pathway

The most common molecular pathway in sporadic CRC (50-70%) is mainly characterised by frequent CIN, MSS, CIMP-N/L and mutations in *APC* and *TP53* (Eshleman *et al.* 1998, Chang *et al.* 2006, Shen *et al.* 2007, Issa 2008). Moreover, *BRAF* is largely unmutated, and the most common *BRAF* mutation (V600E) occurs at a frequency of < 5% in these tumours (Chang *et al.* 2006, Shen *et al.* 2007, Derks *et al.* 2008). Colorectal tumours developing through the CIN positive/MSS pathway are thought to arise from tubular adenomas and occur mainly in the distal colon (Chang *et al.* 2006, Issa 2008). These tumours are associated with an intermediate prognosis and do not show any sex bias (Issa 2008, Ogino and Goel 2008).

1.3.5.1.3 CIMP-Low/MSS

The 3rd main molecular pathway is poorly defined and accounts of 10-30% of sporadic CRC cases (Issa 2008). It is characterised mainly by CIMP-L (lower levels compared to MSI-H/CIMP-H pathway) and mutations in *KRAS* (Ogino *et al.* 2006b, Shen *et al.* 2007). The *BRAF* V600E mutation occurs in some of the cases in which *KRAS* is wild type. This pathway is usually MSS with lower levels of CIN compared to the MSS/CIN positive pathway (Shen *et al.* 2007). Colorectal tumours developing through this pathway are thought to arise from villous adenomas and mainly occur in the proximal colon (Issa 2008). CRC patients with tumours of this type are more commonly males, usually have poor prognosis and their tumours are more resistant to chemotherapeutic agents (Ogino *et al.* 2006b, Shen *et al.* 2007).

It is worth noting that these 3 molecular pathways/subtypes do not cover all the sporadic CRC cases (Issa 2008). Whilst some CRC cases are reported to show one or more of the 3 forms of genomic instabilities, others have been reported without any type of genomic instability (Cheng *et al.* 2008). It is still unclear whether these cases represent separate CRC molecular subtypes, with distinct clinico-pathological features.

1.3.5.2 Standardisation of CRC molecular classification

An important limitation of the molecular classification and characterisation of CRC is the absence of standard techniques and definitions for the global genomic instabilities (Issa 2008, Ogino and Goel 2008). The use of different methods, definitions and marker panels have resulted in major differences in estimating the frequencies of MSI, CIN and CIMP (Issa 2008). In order for molecular classification to generate clinically useful information, standard techniques and definitions must be used to identify the genomic instabilities, especially CIMP and CIN.

1.3.5.3 Sporadic CRC molecular markers and the clinic

1.3.5.3.1 Prognostic markers

A more comprehensive understanding of CRC genetics and epigenetics is paving the way for personalised CRC care and treatment in the clinic. The prognostic values of CIN and MSI were validated in meta-analyses across a wide range of CRC patients (Popat *et al.* 2005, Walther *et al.* 2008, Guastadisegni *et al.* 2010). CIN was associated with a bad prognosis, and was recommended to be evaluated as a prognostic marker in clinical trials (Walther *et al.* 2008, Pritchard and Grady 2011). On the other hand, MSI was associated with a more favourable prognosis (Popat *et al.* 2005, Guastadisegni *et al.* 2010). The strength of these prognostic value means that both CIN and MSI are likely to become increasingly utilised for the clinical evaluation of CRC (Pritchard and Grady 2011). MSI testing is already available for sporadic CRC, however, it is not yet widely used (Pritchard and Grady 2011). Other genetic markers with potential prognostic values include the chromosome 18q LOH and mutations in *BRAF* (V600E), both of which predict an unfavourable prognosis (Popat and Houlston 2005, Roth *et al.* 2010).

1.3.5.3.2 Predictive markers

The most evident example of personalised CRC treatment based on genetic markers is the use of *KRAS* mutational status as a predictive marker for resistance to anti-epidermal growth factor receptor (EGFR) therapies such as cetuximab and panitumumab (Karapetis *et al.* 2008, Allegra *et al.* 2009). Detection of *KRAS* mutations, mostly codons 12 and 13, is currently in routine clinical use, and CRC patients with *KRAS* mutations do not receive anti-EGFR antibodies as part of their treatment (Allegra *et al.* 2009, Pritchard and Grady 2011). In addition to *KRAS* mutations, *BRAF* V600E mutation, *PIK3CA* mutations and loss of PTEN expression are promising predictive markers for resistance to anti-EGFR therapies (particularly in resistant cases with WT *KRAS*) (Figure 1.6) (Di Nicolantonio *et al.* 2008, Jhawer *et al.* 2008, Sartore-Bianchi *et al.* 2009, Di Fiore *et al.* 2010). A combined panel of these mutations might prove to be more efficient in predicting resistance to anti-EGFR therapies (Bohanes *et al.* 2011).

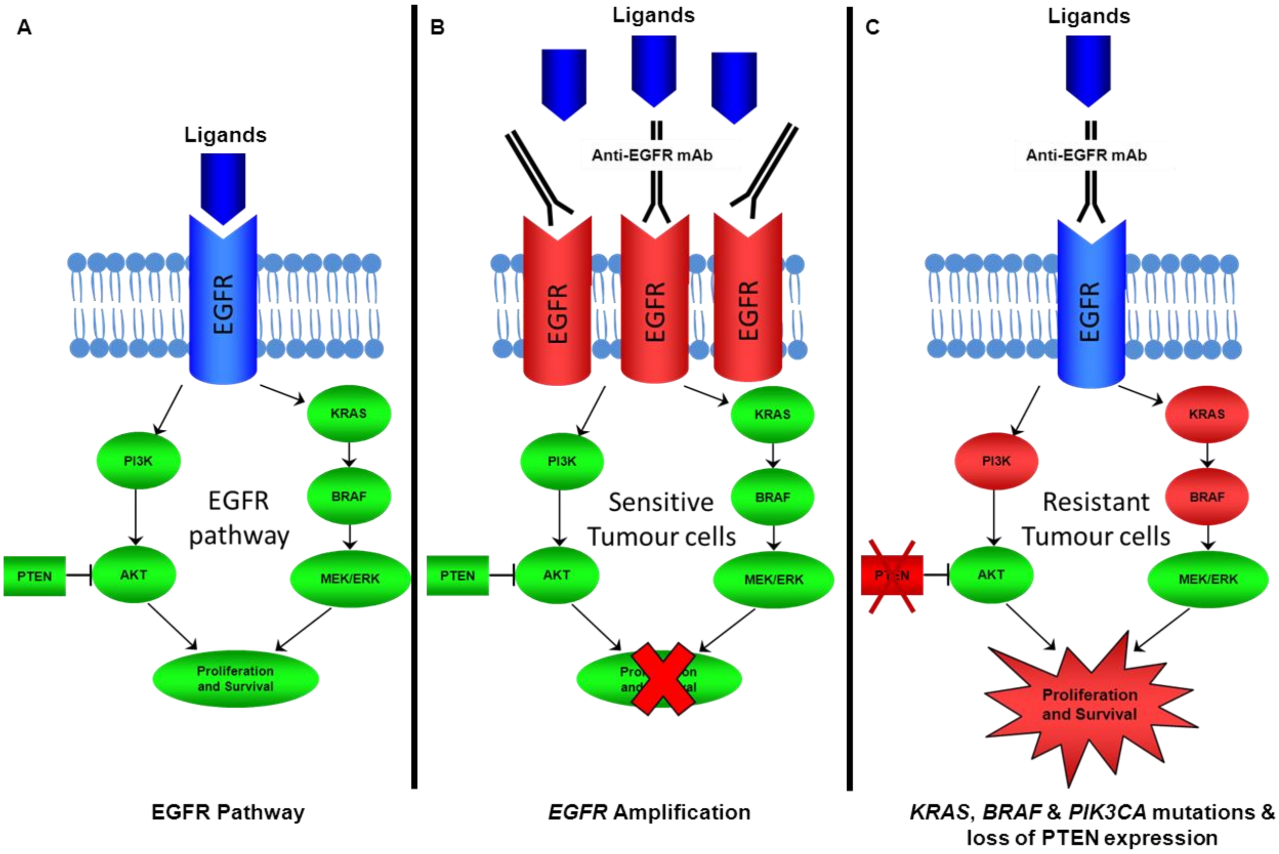


Figure 1.6 Anti-EGFR therapy molecular determinants

A) An overview of the EGFR pathway and the main downstream effector molecules B) In sensitive cells, EGFR activation (overexpression, genomic amplification) can be counteracted by anti-EGFR therapies C) In resistant cells, mutations affecting downstream effector molecules make the treatment ineffective (Adapted from (Di Fiore *et al.* 2010)).

In addition to their prognostic value, MSI and 18q LOH are also promising predictive markers for conventional adjuvant chemotherapy (Pritchard and Grady 2011). MSI was shown to predict resistance to treatment with fluorouracil (5-FU), however, the association is controversial (Ribic *et al.* 2003, Jo and Carethers 2006, Kim *et al.* 2007). A recent meta-analysis of 31 studies has confirmed that the association between MSI and 5-FU resistance is controversial and recommended the use of combined genomic markers to predict the efficiency of 5-FU therapy (Guastadisegni *et al.* 2010). Nevertheless, MSI-H CRC cell lines were shown to be more sensitive to irinotecan treatment (Vilar *et al.* 2008) and a recent clinical trial has shown improved response of MSI patients to adjuvant chemotherapy with irinotecan (Bertagnolli *et al.*

2009). CIMP was also recently shown as a potential predictive marker for resistance to 5-FU adjuvant chemotherapy (Jover *et al.* 2011).

In summary, due to the more comprehensive and detailed understanding of CRC genetic and epigenetic characteristics, personalised CRC treatment is becoming more of a reality (Choong and Tsafnat 2011, Pritchard and Grady 2011). Whilst the use of a single genetic or epigenetic marker is unable to fully predict prognosis or direct treatment decisions, the combinations of several markers through standardised CRC molecular classification provide scope for the future use of personalised CRC treatment (Choong and Tsafnat 2011). Therefore, this demonstrates the need for research directed towards identifying panels of novel genetic and epigenetic markers that are implicated in CRC tumourigenesis.

1.4 Apoptosis

1.4.1 Background

Cell death usually occurs through three main pathways; necrosis, autophagy or apoptosis (Ishimura and Gores 2005). Apoptosis (the Greek word for “falling off”) was described for the first time in 1972 as a mechanism of controlled cell deletion that is complementary but opposite to mitosis (Kerr *et al.* 1972). Apoptosis differs from other pathways of cell death, particularly necrosis, by being highly regulated and occurring via well-defined steps (Okada and Mak 2004). Also referred to as programmed cell death, apoptosis is considered a fundamental and highly controlled biochemical pathway that plays an important role in normal tissue homeostasis, differentiation, immune system function and embryonic development (Ellis *et al.* 1991, Jacobson *et al.* 1997, Reed 2000, Fadeel and Orrenius 2005). However, the deregulation of apoptosis plays a key role in the development of neurodegenerative and autoimmune diseases and cancer (Okada and Mak 2004). Apoptosis is a complicated process and involves the activation and inhibition of several signalling pathways (Huerta *et al.* 2006) (Figure 1.7).

1.4.2 Caspases and apoptosis

The main executioners of apoptosis are a family of proteolytic enzymes called cysteinyl aspartate-specific proteases or caspases (Alnemri *et al.* 1996, Zhao *et al.* 2010). The name “Caspase” reflects the catalytic properties of these proteases; “C” refers to the cysteine protease mechanism, and “aspase” refers to their distinctive ability to cleave their protein targets after an aspartate residue (Alnemri *et al.* 1996).

Up to 14 human caspases have been identified, although not all are implicated in apoptosis (Zhang *et al.* 2004). Caspases that function during apoptosis can be divided into two main groups; initiator or activator caspases characterised by a long N-terminus (Caspases 2, 8, 9 and 10) and executioner caspases characterised by a short N-terminus (Caspases 3, 6 and 7). Caspases are synthesised as inactive zymogens, termed procaspases (Earnshaw *et al.* 1999, Zhang *et al.* 2004). Procaspases can be activated by proteolytic cleavage after an aspartic acid residue, which allows autocatalytic activation of initiator caspases (Muzio *et al.* 1998, Shi 2004). Activated initiator caspases will, in turn, cleave and activate executioner procaspases (Shi 2004). The specificity of the caspases is determined by 4 amino acid residues located near the cleavage site (Thornberry *et al.* 1997). In general, the initiator caspases can activate executioner caspases through 2 classical pathways: the extrinsic death receptor pathway and the intrinsic (mitochondrial) pathway (Figure 1.7) (Zhang *et al.* 2004, Fadeel and Orrenius 2005, Huerta *et al.* 2006).

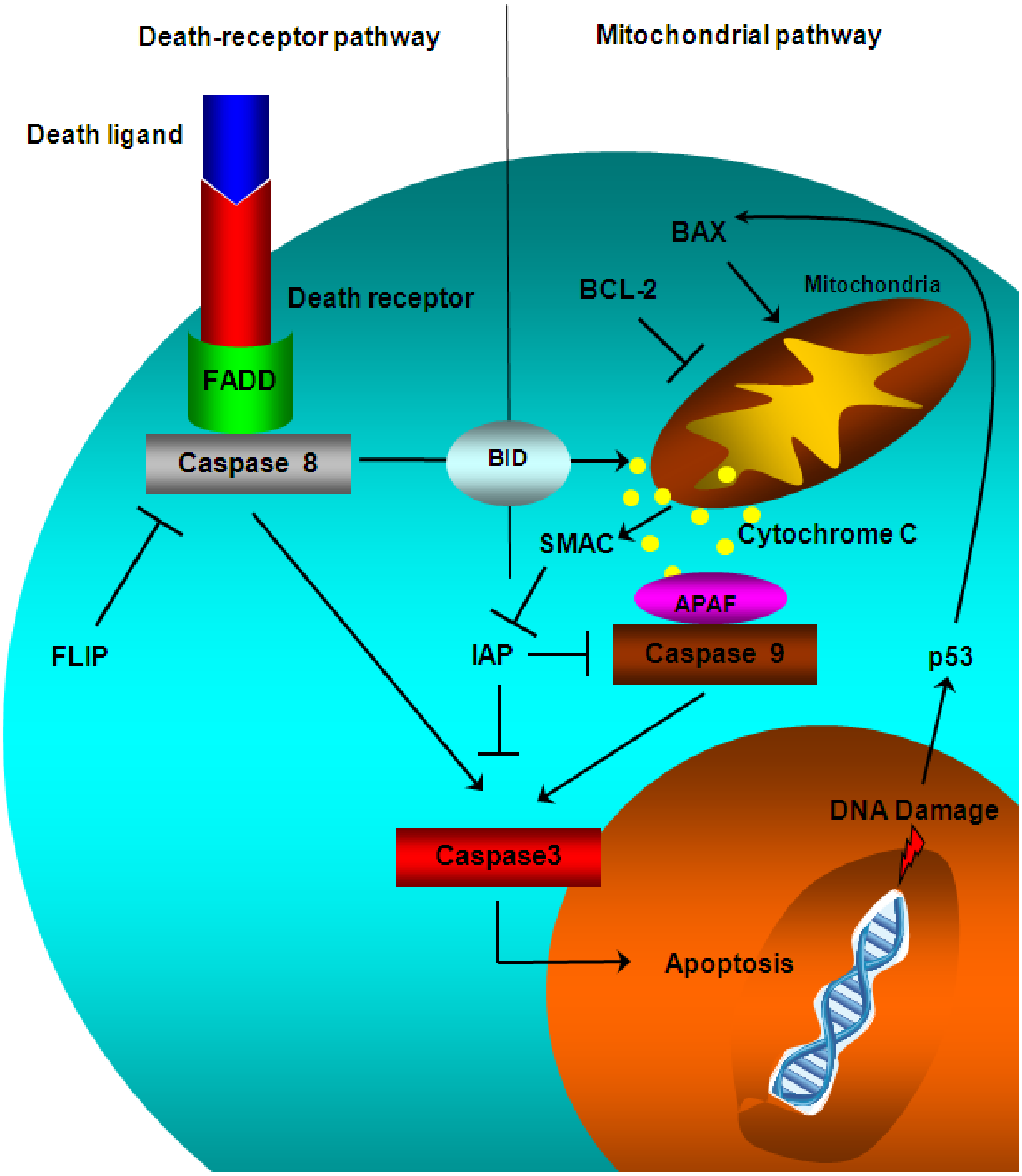


Figure 1.7 **Pathways and regulation of Apoptosis**

Extrinsic (death-receptor) pathway and intrinsic (mitochondrial) pathway. Arrows (→) represent stimulatory effect and dashed lines (⊣) represent inhibitory effects.

1.4.2.1 Pathways of apoptosis

1.4.2.1.1 The extrinsic pathway

The extrinsic pathway is activated through the binding of extracellular ligands such as tumour necrosis factor (TNF) and Fas ligand to various death receptors. The intracellular (cytoplasmic) domains of these death receptors (death domains) subsequently recruit an intracellular adaptor protein known as FADD (Fas-associated death domain protein, also known as MORT1). FADD contains a death effector domain and death domain which binds to the receptor death domain. FADD then attracts and binds procaspase 8 (or procaspase 10) through dimerization of death effector domains (Zhang *et al.* 2004, Elmore 2007). This complex is known as the death-inducing signalling complex (DISC). Procaspase 8 can then be auto-activated through the close proximity of several procaspase 8 molecules (known as the induced proximity model) (Muzio *et al.* 1998, Shi 2004). Active caspase 8 molecules released into the cytoplasm can subsequently cleave and activate executioner caspases. This then results in apoptosis and cell death (Figure 1.7) (Johnstone *et al.* 2002, Huerta *et al.* 2006).

In some cell types, such as lymphoid cells, caspase 8 activation by DISC is sufficient to initiate apoptosis. However, in other cell types, such as hepatocytes, this pathway is not sufficient to fully activate apoptosis. In such cases, the mitochondrial (intrinsic) pathway is required to amplify the “weaker” extrinsic apoptotic signal (Igney and Krammer 2002, Zhang *et al.* 2004). The activation of the mitochondrial pathway in this case is mediated by BID (a pro-apoptotic member of the Bcl-2 family), which is be cleaved and activated by caspase 8 (Figure 1.7) (Luo *et al.* 1998, Yin *et al.* 1999).

1.4.2.1.2 The intrinsic pathway

In the intrinsic pathway, intracellular stress, such as DNA damage, can activate pro-apoptotic proteins of the Bcl-2 family through P53 (Section 1.4.3.1) (Figure 1.7). Pro-apoptotic members of the Bcl-2 family, such as BAX and BAK, stimulate mitochondrial outer membrane permeabilisation and the release of cytochrome C and SMAC/DIABLO (Section 1.4.3.3) (Shimizu *et al.* 1999, Huerta *et al.* 2006). SMAC/DIABLO counteract the anti-apoptotic proteins known as IAP (Section 1.4.3.3), and free cytosolic cytochrome C forms a complex with APAF1 (apoptotic protease activating factor-1) (required for the efficient activation of caspase-9), ATP (adenosine triphosphate) and the inactive form of procaspase-9 (Rodriguez and Lazebnik 1999, Huerta *et al.* 2006). This complex, known as the apoptosome, activates the initiator caspase 9 and results in activation of executioner caspases and apoptosis (Figure 1.7) (Igney and Krammer 2002, Huerta *et al.* 2006).

Following the activation of either the extrinsic or intrinsic pathway, activated executioner caspases (mainly caspases 3, 6 and 7) can cleave specific “death substrates”, leading to morphological changes including DNA fragmentation, chromatin condensation and cytoplasmic shrinkage, apoptosis and cellular death (Zhang *et al.* 2004, Huerta *et al.* 2006).

1.4.3 Regulation of apoptosis

Apoptosis is a highly controlled process, and several protein families including B cell leukaemia/lymphoma 2 (Bcl-2) family proteins, inhibitors of apoptosis proteins (IAP), CASP8 and FADD-like apoptosis regulator (CFLAR) and heat shock proteins (HSP) play a role in its regulation. Interactions between these proteins and the apoptotic machinery determine the activation or the inhibition of apoptosis at different levels (Igney and Krammer 2002, Zhang *et al.* 2004).

1.4.3.1 Bcl-2

The Bcl-2 protein family regulate apoptosis at the mitochondrial level (intrinsic pathway). They are divided into pro-apoptotic proteins such as Bcl-2–associated X protein (BAX), BCL-2 antagonist/killer (BAK) and BH3-interacting domain death agonist (BID) and anti-apoptotic proteins such as Bcl-2, Bcl-XL and Bcl-W (Figure

1.7). Bcl-2 proteins regulate the intrinsic pathway by altering the permeability of the mitochondrial membrane (Igney and Krammer 2002, Huerta *et al.* 2006).

1.4.3.2 CFLAR (FLIP)

By interfering directly with DISC, FLIP inhibits the initiation of apoptosis via the extrinsic pathway. There are 2 splice variants of FLIP; long variant (FLIP$_L$) and short variant (FLIP$_S$). Both forms bind to DISC and inhibit the activation of caspase 8 and the extrinsic pathway (Figure 1.7) (Igney and Krammer 2002).

1.4.3.3 IAP

Nine IAP family members, including XIAP, apollon and survivin, are known to inhibit apoptosis by directly interfering with and inhibiting caspases 3, 7 and 9 (Igney and Krammer 2002, Zhang *et al.* 2004). During apoptosis, a protein known as second mitochondria-derived activator of caspase/direct IAP binding protein with low pI (SMAC/DIABLO), is released from the mitochondria. This binds and inhibits IAPs in order to allow the activation of caspases (Figure 1.7) (Igney and Krammer 2002).

1.4.3.4 HSP

HSP influence both the intrinsic and extrinsic pathways of apoptosis (Huerta *et al.* 2006). Anti-apoptotic HSP such as HSP27 inhibit the release of cytochrome C, while HSP70 and HSP90 inhibit the apoptosome through interfering with APAF-1. On the other hand, pro-apoptotic HSP such as HSP10 and HSP60 stimulate caspase 3 (Huerta *et al.* 2006).

1.4.4 Apoptosis and colorectal cancer

Deregulation or inhibition of apoptosis has been shown to play an important role in the initiation, development and progression of cancer (Lowe and Lin 2000, Johnstone *et al.* 2002). In order to proliferate indefinitely, tumour cells must be able to avoid apoptosis (Hanahan and Weinberg 2000). It was previously shown that apoptosis is inhibited or downregulated during the development of CRC (Bedi *et al.* 1995, Kawasaki *et al.* 1998). The disruption of the apoptotic pathway usually occurs through the abnormal expression of genes that control and regulate apoptosis (Butler *et al.* 1999). This generally arises through loss of function in genes that upregulate or activate apoptosis (such as *TP53* and *BAX*) and overexpression of genes that inhibit or downregulate apoptosis (such as *BCL2*, *BIRC5* and *FLIP*) (Kawasaki *et al.* 1998, Butler *et al.* 1999, Ryu *et al.* 2001, Huerta *et al.* 2006). Besides uncontrolled growth, defective apoptosis helps tumour cells to circumvent the body's immune defences (Ishimura and Gores 2005).

1.4.5 Apoptosis and chemotherapy

Cancer treatments – including chemotherapy and irradiation – mainly act through the induction of apoptosis. Therefore, defects in the apoptotic machinery usually result in cancer cells exhibiting resistance to treatment (Ishimura and Gores 2005). For example, mutations in the pro-apoptotic *BAX* gene were shown to render CRC cells resistant to treatment with nonsteroidal anti-inflammatory drugs (NSAID) (Geelen *et al.* 2004, Zhang *et al.* 2004). On the other hand, overexpression of the anti-apoptotic protein FLIP resulted in resistance of CRC cell lines to treatment with CRC drugs such as 5-FU, OXA (oxaliplatin) and CPT-11 (irinotecan) (Geelen *et al.* 2004, Longley *et al.* 2006).

1.4.6 Apoptosis and genomic instability

Apoptosis and genomic instability are closely linked in tumourigenesis (Zhivotovsky and Kroemer 2004). Genomic instabilities usually result in mutations or abnormal expression of genes that regulate apoptosis. For example, MSI often results in mutations that inactivate the pro-apoptotic genes *BAX*, *CASP5* and *BCL10* (Trojan *et al.* 2004, Iacopetta *et al.* 2010). Moreover, deletion of the short arm of chromosome 17 (17p), as a result of CIN, usually leads to *TP53* inactivation (Zhivotovsky and Kroemer 2004, Ozakyol *et al.* 2006). P53 protein plays a central role in initiating apoptosis, keeping the integrity of the genome and thus preventing tumourigenesis (Bourdon 2007, Vazquez *et al.* 2008). In fact, the *TP53* gene is inactivated in up to

85% of human cancers (Huerta *et al.* 2006). Simultaneously, defective apoptosis enables the survival and division of cells with unstable genomes, which otherwise would be eliminated (Zhivotovsky and Kroemer 2004).

1.4.7 Caspase 8

As mentioned earlier, caspase 8 is an initiator caspase that plays a key role in the activation of the extrinsic pathway. Moreover, through its cleavage of BID, caspase 8 also plays an important role in the activation of the intrinsic pathway (Igney and Krammer 2002). In 1999, the caspase 8 gene (*CASP8*) was mapped to chr2q33 downstream of *CFLAR* and *CASP10* and upstream of *ALS2CR12*. *CASP8* is approximately 54Kb in size and contains 13 exons (Figure 1.8) (Grenet *et al.* 1999). There are 8 identified isoforms of caspase 8 (a-h), however, only isoforms, a and b are predominantly expressed (Scaffidi *et al.* 1997), and they result in proteins containing 496 and 479 amino acids respectively (relative molecular mass of 55.4KDa (isoform a) and 53.7KDa (isoform b).

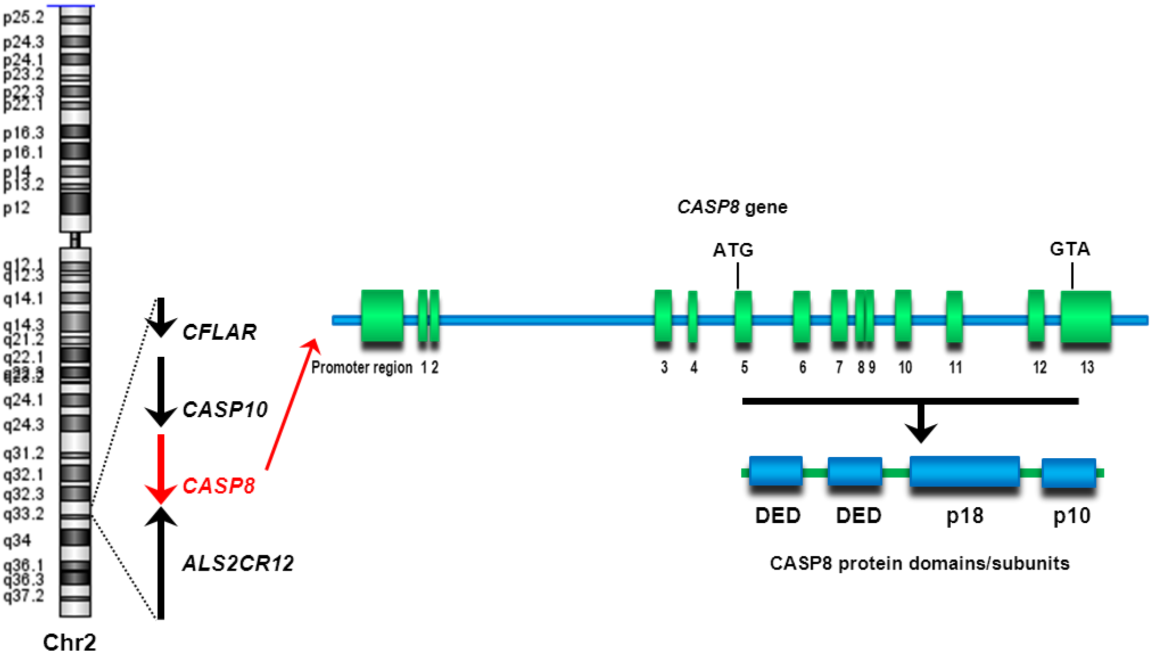


Figure 1.8 ***CASP8*** **genomic location, gene structure & protein**

1.4.7.1 Caspase 8 in cancer and CRC

Deletion of chromosome 2q33, the genomic region containing *CASP8*, has been reported in several cancers including gastric and lung cancer and neuroblastoma (Otsuka *et al.* 1996, Nishizuka *et al.* 1998, Teitz *et al.* 2000, Takita *et al.* 2001, Shivapurkar *et al.* 2002a, Geelen *et al.* 2004). Moreover, it was reported that *CASP8* was frequently silenced by methylation in small cell lung cancer (SCLC) cell lines, neuroblastoma and medulloblastoma (Teitz *et al.* 2000, Shivapurkar *et al.* 2002a, Zuzak *et al.* 2002, Pingoud-Meier *et al.* 2003a, Stupack *et al.* 2006). Furthermore, *CASP8* somatic mutations were reported in ~11% of advanced gastric cancers (Soung *et al.* 2005) and >5% of advanced CRC cases (Kim *et al.* 2003). Loss of *CASP8* (by inactivating mutations, genomic deletions or methylation) plays a role in inducing resistance to chemotherapy, in promoting metastasis and is associated with poor prognosis and survival (Hopkins-Donaldson *et al.* 2000, Kim *et al.* 2003, Pingoud-Meier *et al.* 2003a, Stupack *et al.* 2006)

1.4.7.2 *CASP8* polymorphisms and cancer risk

Common polymorphisms in *CASP8* were shown to affect lung cancer risk (Son *et al.* 2006). Moreover, a common missense variant (D302H) in exon 12 of *CASP8* was reported as a low penetrance susceptibility allele for breast cancer (MacPherson *et al.* 2004, Cox *et al.* 2007, Shephard *et al.* 2009). In addition, a 6bp insertion-deletion (in/del) polymorphism in the promoter region of *CASP8* was shown to affect the susceptibility of several common cancers including CRC (Sun *et al.* 2007). However, results of the latter association were not confirmed by additional studies (Haiman *et al.* 2008, Pittman *et al.* 2008). Nevertheless, preliminary data from our laboratory have indicated a possible association between *CASP8* SNPs and CRC risk (Curtin *et al.* manuscript in preparation).

1.5 Aims of the project

Genetic factors (both inherited and somatic) play a key role in the predisposition, initiation and development of CRC. Large family based cancer studies involving 1st degree relatives and twin siblings have estimated that the genetic contribution towards CRC risk is very high (~35%) (Goldgar *et al.* 1994, Lichtenstein *et al.* 2000, Dong and Hemminki 2001, de la Chapelle 2004). Whilst some of the inherited factors are highly penetrant and rare, others are of low penetrance and are more common. Highly penetrant CRC genetic factors are well characterised (e.g. Lynch and FAP syndromes), however, they account for <20% of hereditary CRC (de la Chapelle 2004, Migliore *et al.* 2011). The remaining cases of hereditary CRC are probably due to the more common, but less penetrant, genetic polymorphisms. Several low penetrant and common CRC risk variants have been identified through genome wide and candidate genes association studies, however, so far they only account for ~6% of CRC heritability (Lascorz *et al.* 2010). By identifying more of these common and low penetrant genetic factors, further biochemical pathways involved in the pathogenesis of CRC might be identified, with the potential of extensive clinical implications (de la Chapelle 2004).

Cellular evasion of apoptosis is considered one of the hallmarks of cancer development (Hanahan and Weinberg 2011). This is usually achieved by the disruption of genes which regulate apoptosis (Butler *et al.* 1999). Caspase 8 is an initiator caspase with an important role in the regulation of apoptosis. The mutation or methylation of *CASP8* is implicated in several cancers including CRC (Teitz *et al.* 2000, Kim *et al.* 2003, Soung *et al.* 2005). Moreover, inherited *CASP8* variants were shown to affect cancer risk, including CRC (Son *et al.* 2006, Cox *et al.* 2007, Sun *et al.* 2007). However, the association between *CASP8* inherited variants and CRC risk is controversial (Haiman *et al.* 2008, Pittman *et al.* 2008). The aims of the first part of the project were to further test the hypothesis that *CASP8* inherited variants may be involved in CRC risk. This included the following:

1. To investigate the controversial association between *CASP8* in/del promoter variant and CRC risk by genotyping the in/del variant in 1193 CRC cases and 1388 matching controls.
2. To identify any novel or rare *CASP8* variants with the potential to affect CRC risk by the sequencing of the *CASP8* promoter region, exons, exon/intron boundaries and the 3'UTR in 94 CRC cases.
3. To develop an assay for the *CASP8* copy number variant CNV23598, located close to the D302H (breast cancer risk) variant in a panel of 284 CRC cases and controls, and in 47 carefully selected breast cancer cases and controls enriched for *CASP8* risk and protective haplotypes respectively.

Sporadic CRC develops through different pathways of genetic and epigenetic defects and somatic mutations in key genes. Due to its potential clinical implications, there is increased recognition and interest in the molecular classification of CRC. However, the field is limited by the lack of standardisation in defining global genomic instabilities (CIN, MSI and CIMP) (Issa 2008, Ogino and Goel 2008). Therefore, the main aim of the 2nd part of the project was to molecularly classify 53 CRC cases using standard definitions and technologies. This involved the following:

1. To investigate and define MSI status using a panel of 5 mononucleotide microsatellite markers according to the revised NCI criteria for MSI testing.
2. To investigate CIMP using the latest validated panel of 7 methylation markers and the semi-quantitative methylation-specific multiplex ligation probe amplification (MS-MLPA) technology.
3. To investigate CIN using high resolution genome wide array CGH.

CIN is considered the most common genomic instability in CRC. CIN plays a major role in CRC development through the amplification of oncogenes (such as *MYC*) and the deletion of TSG (such as *SMAD4*) (Migliore *et al.* 2011). However, the picture is far from complete (Brosens *et al.* 2010, Pino and Chung 2010). Therefore, the final aim of the project was to identify and characterise novel CRC driver genes affected by CIN. This involved:

1. The identification of common focal chromosomal aberrations based on genome wide high resolution aCGH data.
2. Proof of concept experiments to examine the effects of alterations on one of the novel candidate driver genes.

2. MATERIALS AND METHODS

2.1 Materials

2.1.1 General laboratory equipment and consumables

Table 2.1 Laboratory equipments

Equipment	Supplier
ABI 7900 Genotyping Platform	Applied Biosystems
Agarose Gel Electrophoresis Unit (Sub-cell GT)	Bio-Rad
AB104-S Balance	Mettler, Toledo
Class II Microbiological Safety Cabinet	Envair
CO_2 Incubator MCO175	Sanyo
Fujifilm LAS3000 Chemiluminescence Imaging system	Fujifilm
GeneAmp® PCR System 2700 (9700) Thermal Cycler 96 well	Applied Biosystems
Heating Block	Grant Boekel BBA
Heated Plate Sealer ALPS^{TM} 50	Abgene
Heraeus Pico 17 Centrifuge	Thermo Scientific
Harrier 15/80-MSE Centrifuge	Sanyo
Ice Machine	Scotsman Ice Machines
Impact 2 Multichannel Electronic Pipette (0.5-10ml)	Matrix
Multichannel Pipette (0.5-10ml)	Matrix
Multiskan FC spoectrophotometer	Thermo Scientific
Nanodrop	Thermo Scientific
P2, P10, P20, P200, P1000 Pipettes	Gilson
Peltier Thermal Cycler (DNA Engine Dyad ^{TM}) 384 well	MJresearch Inc
Power Pack	Bio-Rad
Powerpette Plus Pipette	Jencons
QBD4 Incubator for eppendorfs	Grant
Semi-Dry transfer Cell	Bio-Rad
Titramax 1000 Incubator and Shaker	Heidolph
U:Genius Gel Imager	Syngene
UV Sterilisation Cabinet	Bigneat
Vortex Genie 2	Scientific Industries
Western Gel Mini Protean II Cell	Bio-Rad

Table 2.2 Laboratory consumables

Equipment	Supplier
10ml and 25ml Stripettes	Corning Inc
15ml Centrifuge Tubes	Sarstedt
6 and 96 Well Tissue Culture Plates	Greiner Bio-One, Corning Inc
Cell culture petri dishes	Greiner Bio-One
384 Well PCR Plates	Applied Biosystems
50ml Centrifuge Tubes	Corning Inc
500µl, 1.5ml and 2ml Microfuge Tubes	Sarstedt
96 Well PCR Plates	Starlab
Pipette Tips	Matrix, Bioscience, Starlab, Axygen
Gloves	Schottlander
Heat Sealing Foil	Thermo Scientific
Microseal B Adhesive Sealer	Bio-Rad
Pasteur Pipettes	Scientific Laboratory Supplies
Plate Seals	Alphalaboratories
Reservoirs with Dividers	Thermo Scientific

2.1.2 Laboratory solutions

- TAE buffer (1 L, 10x, pH8.0): 0.4 M Tris-base, 11 ml glacial acetic acid, 10 mM EDTA and ddH_2O (pH adjusted to 8.0).
- EDTA (1 L, 0.5 M, pH8.0): 1.861g sodium ethylene diamine tetra-acetate in 1 L ddH_2O (pH adjusted to 8.0).
- Phosphate Buffered Saline (PBS) (100 ml, 1x): 1 Oxoid PBS tablet in 100 ml ddH_2O.
- Transfer buffer (1 L, 10x): 30.3 g (0.25 M) Tris base, 144 g (1.92 M) glycine and ddH_2O.
- TBS (1 L, 10x): 87.66 g (1.56 M) NaCl, 24.22 g (0.2 M) Tris and ddH_2O.
- SDS-PAGE running buffer: 30.3 g (0.25 M) Tris base, 144 g (1.92 M) glycine, 10 g (1%) Sodium Dodecyl Sulphate (SDS) and ddH_2O.
- Tris-Cl (1 L, 1 M, pH6.8 or pH8.8): 121.1 g Tris-Cl in ddH_2O (pH adjusted to 6.8 or 8.8 as needed).

2.1.3 DNA samples

2.1.3.1 Sheffield samples

Peripheral blood DNA samples from Sheffield and the North Trent region were available from patients who had colorectal tumour surgery at the Sheffield Royal Hallamshire, Sheffield Northern General, Chesterfield Royal Infirmary, Doncaster Royal Infirmary and Barnsley District General hospitals (March, 2001 – June, 2005). These samples were identified with a 3 digit number prefixed with a "C". Peripheral blood DNA samples were also available from patients who had chemotherapy treatment for metastatic CRC at Sheffield Weston Park hospital (March, 2001 – June, 2005). These samples are referred to as "metastatic" and the samples numbers are prefixed with an "M". Age- and sex- matched controls were recruited through the general practice (GP) registers in Sheffield (October, 2001 – December, 2005) and DNA samples were also available from their peripheral blood samples.

2.1.3.2 Leeds and Dundee samples

Peripheral blood DNA Samples from Leeds and Dundee were available from patients taking part in a population-based CRC epidemiological study in the North of England (Barrett *et al.* 2003). In Leeds, the cases were identified and recruited in Leeds General Infirmary and St James's hospitals. In Dundee, the cases were recruited from Ninewells and Perth Royal Infirmary Hospitals (1997 – 2000). Peripheral blood DNA samples from age- and sex- matched controls were also available from both Leeds and Dundee GP registers.

2.1.3.3 Utah samples

In Utah, peripheral blood DNA samples were extracted from CRC cases selected from high risk cancer pedigrees listed in the Utah Cancer Registry (UCR) and Utah Population DataBase (UPDB). Peripheral blood DNA samples were also available from cancer-free controls recruited for different studies. Controls were matched based on sex and year of birth (within a 5 year age range of the cases).

2.1.3.4 Paired tumour and normal DNA samples

For 50 patients from the Sheffield population sample, DNA samples were also available from freshly frozen tumour tissue in addition to the peripheral blood. These DNA samples were extracted from micro-dissected colorectal tumour tissues that contained at least 80% cancerous cells. Throughout this thesis, this cohort of matching DNA samples will be referred to as the "Sheffield paired cohort" and the DNA extracted from the peripheral blood and tumour tissue will be referred to as "normal" DNA and "tumour" DNA, respectively. Patients from this cohort will be identified with a 3 digit number prefixed with "CA" and the normal and tumour DNA samples for each patient will have the same number prefixed with "N" and "T" respectively. Clinical data including gender, diagnosis age, tumour location and stage, histological type, tumour differentiation and presence of metastasis were available. Follow up data on patient survival are available for these patients until March 2011. These data were obtained from the Trent cancer registry.

Seventeen additional paired DNA samples were available from the tissue bank in the Royal Hallamshire hospital- University of Sheffield. For these samples, normal DNA was extracted from freshly frozen normal colon tissues taken at the time of surgery. All tumour and normal tissue samples were checked independently by the pathologist Jonathan Bury before being included in the study. Clinical data were also available for these patients, however, no follow up information was available. These tissue bank patients and their DNA samples will have the same numbering identifiers explained above for the Sheffield paired cohort.

All the subjects included in the study have given informed consent for data and sample collection and ethical approval of the study was obtained from South Yorkshire Research Ethics Committee (Appendix 3).

The above DNA sample sets were previously used in published genetic epidemiological studies (Curtin *et al.* 2009a, Curtin *et al.* 2009b).

2.1.4 Cell lines

The colon adenocarcinoma cell lines (HCT116, SW620, SW480, HT29, CACO2 and COLO205), the Glioblastoma Multiforme (GBM) cell lines (U87 and SNB19) and the non-tumour derived cell lines HEK293 (human embryonic kidney epithelial) and MRC5-SV2 (foetal lung fibroblast) were supplied by the American Type Culture Collection (ATCC), (Bethesda, USA).

2.1.5 PCR and sequencing primers

Primer design is described in Section 2.2.5. Primer sequences are provided in Appendix 1, tables 1-7. Primers used in this project were obtained from Sigma Aldrich (Ebersberg, Germany). The primers were supplied lyophilised and were reconstituted with deionised water at a stock concentration of 100 pmol/ µl (as per supplier’s recommendations).

2.1.6 Pyrosequencing primers

Pyrosequencing primer design is described in Section 2.2.14.1. Primer sequences are provided in Appendix 1, table 8. Pyrosequencing primers were ordered from Applied Biosystems (Warrington, UK). The primers were supplied lyophilised and were reconstituted with deionised water at a stock concentration of 100 pmol/ µl (as per supplier’s recommendations).

2.2 Methods

2.2.1 DNA extraction

2.2.1.1 Cultured cells

DNA from cancer cell lines was extracted using QIAamp DNA mini kit (Qiagen, West Sussex) according to the manufacturer's instructions. Briefly, the cells were split into six 1.5ml tubes and centrifuged at 300 *x*g for 5 min. The supernatant was removed and the cell pellet was re-suspended in 200 µl 1x phosphate buffer saline (PBS). The cells were then lysed by the addition of 20 µl proteinase K, 200 µl buffer AL, and incubation at 56^{o}C for 10 min. The tubes were briefly centrifuged and 200 µl ethanol was added. Next, this mixture was transferred to QIAamp mini spin columns (provided in the kit) and centrifuged at 6000 *x*g for 1 min. The pellets were then washed twice by adding the provided wash buffers 1 & 2 with centrifugation at 6000 *x*g for 1 min each time. An extra centrifugation step (13000 *x*g, 3 min) was performed to remove any remains of wash buffer 2. Finally, the DNA was eluted by the addition of 200 µl elution buffer, 1 min incubation at room temperature, followed by 1 min centrifugation at 6000 *x*g.

2.2.1.2 Frozen tissues

Frozen tissue DNA was extracted using the QIAamp DNA mini kit (Qiagen) according to the manufacturer's instructions. Briefly, the colon tissues were cut into small pieces using a scalpel. The pieces were then transferred to a 1.5 ml tube and 180 µl of buffer ATL and 20 µl of proteinase K were added. The samples were then mixed by vortexing and incubated at 56^{o}C for 15 min, followed by manual homogenisation and incubation in an orbital shaker at 56^{o}C overnight. The next day, the tubes were centrifuged and 4 µl of RNase A (100 mg/ml) were added. After pulse vortexing for 15 sec, the samples were left at room temperature for 2 min. Next, 200 µl of buffer

AL were added, the samples mixed by vortexing and then incubated at 70^{o}C for 10 min. The remaining steps were carried out as described in Section 2.2.1.1.

2.2.2 DNA quantification and quality analysis

DNA samples were quantified using a NanoDrop spectrophotometer (ND-1000) (Fisher Scientific UK Ltd., Loughborough). The ND-1000 software was used with the nucleic acids module being selected. After cleaning the sensors, the NanoDrop was initialised and blanked using appropriate blank (1x Tris EDTA (TE), water or Qiagen elution buffer). Then each sample was quantified by pipetting 1.2 µl onto the sensors. The sensors were wiped between different measurements. DNA concentration was calculated using a modified Beer-Lambert equation (c = (A * e) /b) where c is DNA concentration (ng/ µl), A is absorbance at 260nm, e is the absorbance coefficient and equals 50ng-cm/ µl and b is the path length in cm. The purity of the DNA samples was determined using A_{260}/A_{280} and A_{260}/A_{230} ratios. The former indicates the absence of contaminating proteins when the ratio is 1.8-2.0 and the latter indicates absence of other organic compounds when the ratio is 1.8-2.2.

2.2.3 Tumour DNA fragmentation

In order to assess their degree of degradation, tumour DNA samples were electrophoresed on 1% Agarose gel (Section 2.2.7). Prior to loading, 1 µl of each sample was mixed with 3 µl water and 1 µl DNA loading buffer (6x) (0.25% (w/v) bromophenol blue, 30% (v/v) glycerol).

2.2.4 Sodium acetate/ethanol DNA precipitation

Some of the DNA samples used in the project had a low 260/230 ratio which indicates contamination with organic substances such as phenol. Such contamination can inhibit some of the tests applied to the DNA samples. In order to remove the contamination, sodium acetate (NaAc) / ethanol purification was performed. The procedure involved adding a one tenth volume of 3 M NaAc to the DNA sample followed by the addition of 2 volumes of 100% ice cold ethanol. This mixture was then incubated at -20^{o}C for 90 min. The samples were then centrifuged at 14100 RPM for 15 min at 4^{o}C and the supernatant was removed. The samples were spun again and any residual supernatant was removed. Then, 200 µl of 70% ice cold ethanol were added to the pellets which were centrifuged at 14100 RPM for 5 min at 4^{o}C. As much as possible of the supernatant was removed and the remainder was left to evaporate in a laminar flowhood for 15 min. DNA was reconstituted by the addition of 100 µl of 1x molecular grade TE (Promega, Southampton) and the 260/230 ratio was checked again using the NanoDrop (Section 2.2.2).

2.2.5 Primer design

2.2.5.1 *In silico* primer design

Primers used in the amplification of *CASP8*, *APC*, *TP53*, *KRAS*, *BRAF*, *PIK3CA* and *CASP8* CNV23598 were designed using the web-based tool Primer 3 (http://frodo.wi.mit.edu/, accessed October, 2008 – August, 2010) (Rozen and Skaletsky 2000). The target sequence for each region was obtained from NCBI (www.ncbi.nlm.nih.gov, accessed October 2008 – August, 2010) (Table 2.3). The sequences were then imported into the Primer3 web page. The default parameters for picking primers were maintained except for the melting temperature (T_M); primer T_M range of 57°C-63°C was used with 60°C selected as the optimal temperature. For primer size, a range between 20-24 nucleotides was used with 22 as the optimal size and for primer GC%, a range of 45-55% was set with 50% as the optimal GC content. Primer3 provided a list of suggested primers according to the parameters chosen and indicated the most suitable pair. The "best pair" was chosen in all cases. The primer blast search tool provided by NCBI (http://www.ncbi.nlm.nih.gov/tools/primer-blast/, accessed February, 2009 – August, 2010) was then used to test whether the primers were primarily/solely specific to their targets. Moreover, all the designed primers were tested for the occurrence of SNPs underneath their annealing sites using the diagnostic SNP check online tool provided by the National Genetics Reference Laboratories (NGRL) (http://ngrl.man.ac.uk/SNPCheck/SNPCheck.html, accessed October, 2008 – August, 2010).

Table 2.3 NCBI accession numbers

Gene	Chromosome	Accession number	Location	Genomic Build
CASP8	2q33-q34	NC_000002.10	201804871-201862219	36.3
APC	5q21-q22	NC_000005.8	112100455-112210835	36.3
TP53	17p13.1	NC_000017.9	7532642-7511445	36.3
KRAS	12p12.1	NG_007524.1	5001-50675	36.3
BRAF	7q34	NG_007873.1	5001-195753	36.3
PIK3CA	3q26.3	NG_012113.1	5001-91190	37.1

2.2.5.2 Manual primer design

In a few cases, manual primer design was required to target a very specific area. The primers were designed manually to meet the following criteria:

- 20-25 bases in length
- GC content of 45-55%
- 3' end to terminate with G or C
- Single base repeats >3 were avoided
- Minimum complementarities between pairs of F and R primers
- T_M in the range 59ºC - 61ºC. Primer T_M was calculated using the following equation (as recommended by MWG Biotech Ltd.):

$$T_M=69.3 + (GC\% \times 0.41)-(650/\text{primer length}).$$

M13 tails (Table 2.4) were attached to all the primers used in this project, except for long range PCR primers, internal sequencing primers and fluorescently labelled primers. M13 primers are universal primers used for DNA sequencing. They were used to facilitate the sequencing procedure, as all the exons can be sequenced using a single primer pair.

Table 2.4 M13 sequences

Primer	Sequence (5'-3')
M13 Forward	dTGTAAAACGACGGCCAGT
M13 Reverse	dCAGGAAACAGCTATGACC

2.2.5.3 DNA sequencing primer design

Conventional Sanger sequencing techniques are able to analyse PCR products up to 600-800 nucleotides in length. The long-range PCR products (Section 2.2.6.2) in this study ranged between 1000bp up to 3000bp in length. To sequence across these regions, sequencing primers were required. These primers were designed ~400 bp apart to ensure enough overlap between the sequences to confidently cover the region of interest. Primers were designed manually using the same criteria as in Section 2.2.5.2.

2.2.6 Polymerase Chain Reaction

2.2.6.1 Conventional PCR

Exons that required mutation screening were amplified using the polymerase chain reaction (PCR) (Mullis *et al.* 1986). Standard 25 µl PCR reactions were used. Each reaction contained 25 ng genomic DNA, 100 ng of each specific primer (forward and reverse), 1 µl deionised water and 22.5 µl Reddymix® (1.25 U Taq polymerase, 75 mM Tris-HCl, 20mM $(NH_4)_2SO_4$, 0.01% (w/v) Tween20, 0.2 mM of each dNTP and 1.5 mM $MgCl_2$ (Fisher Scientific UK Ltd.). A conventional PCR includes cycles of 3 essential steps: denaturation, annealing and extension (Figure 2.1). The cycling conditions used in this study were; initial denaturation step at 94ºC for 7 min, followed by 35 cycles of denaturation at 94ºC for 1 min, annealing at nºC for 1 min and extension at 72ºC for 1 min, and a final extension step at 72 ºC for 7 min (Figure

2.1). The annealing temperatures varied according to the T_M of the primers (Appendix 1). Throughout the project, all PCR reactions were carried out using the GeneAmp® PCR System 9200 thermocycler (Applied Biosystems).

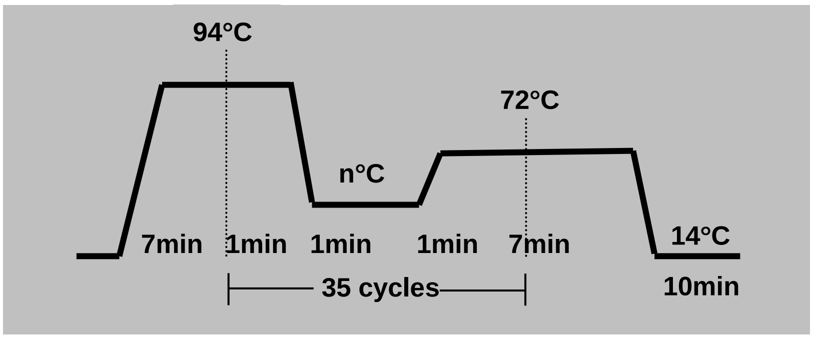


Figure 2.1 Conventional PCR thermal profile
35 cycles of: denaturation (at 94ºC), annealing (depends on primers T_MºC), extension (at 72ºC)

2.2.6.2 Long-range PCR

Long-range PCR is used to amplify larger fragments of DNA than those achieved using conventional PCR protocols (Cheng *et al.* 1994). Modifications from the conventional PCR protocol included the use of thermostable DNA polymerases with high 5'-3' polymerase activity and 3'-5' proofreading capability. Additionally, the extension step in the long-range PCR protocol was usually longer than in conventional protocols to allow the amplification of large fragments. Long-range PCR was performed using the Phusion high fidelity PCR kit (New England Biolabs (UK) Ltd., Hitchin). Long-range PCR reactions were set up in a total volume of 50 µl and consisted of 62.5 ng genomic DNA, 0.5 µM of each primer (forward and reverse), 32.5 µl deionised water and 10 µl of 5x Phusion HF buffer®, 200 µM of each dNTP and 10 ul of the Phusion DNA polymerase. The long-range PCR conditions consisted of the following steps; initial denaturation at 98ºC for 30 sec, followed by 35 cycles of denaturation at 98 ºC for 10 sec, annealing at nºC for 30 sec and extension at 72 ºC for x min, and a final extension at 72ºC for 10 min. The annealing temperatures varied according to the T_M of the primers (Appendix 1) and the extension time varied according to the size of the target region (15-30 sec/Kb). The PCR conditions were determined according to the Phusion kit recommendations.

2.2.6.3 GAP PCR

GAP PCR is a modified PCR that allows the detection of known insertions and deletions. It involves the use of 3 primers: a "common" forward primer, which anneals to the sequence upstream of the insertion/deletion site and 2 reverse primers. The first primer "insertion" anneals to the inserted sequence when present, and the 2nd primer "deletion" will anneal downstream of the insertion/deletion sequence (Figure 2.2). In the case of a homozygous deletion, the "common" and the "deletion" primers alone will produce a product as the "insertion" primer will fail to anneal. In the case of an insertion, the "common" and the "insertion" primers alone will produce a PCR product as the optimised PCR conditions do not allow the amplification of large PCR products. The genotype of the sample is determined by gel electrophoresis based on the size of the different products (Figure 2.2). In this project GAP PCR was initially used for genotyping the CNV23598. Standard conventional PCR reactions and conditions were used, however, with 3 primers (Section 2.2.6.1).

2.2.6.4 Fluorescent GAP PCR

Fluorescent GAP PCR was used as a higher throughput method for the genotyping of *CASP8* CNV23598. The principle is the same as described in the previous section. However, the "insertion" and the "deletion" primers were labelled with distinct fluorescent dyes and instead of gel based analysis, capillary electrophoresis was used. The products appear as peaks of different size and fluorescent colour according to the different alleles. Fluorescent GAP PCR was performed using the Amplitaq Gold® PCR master mix (Applied Biosystems). The reactions were set up in a 10 µl volume and consisted of 5 ng DNA, 0.5 µM of the unlabelled common forward primer, 0.5 µM of the “insertion” labelled reverse primer, 0.2 µM of the “deletion” labelled reverse primer, 5 µl of the Amplitaq Gold® PCR master mix and 2.8 µl deionised water. The PCR conditions consisted of the following steps; initial denaturation at 95ºC for 5 min, followed by 32 cycles of denaturation at 95 ºC for 30 sec, annealing at 59ºC for 30 sec and extension at 72 ºC for 30 sec, and a final extension at 72ºC for 7 min. The products were diluted 1 in 10 with deionised water and then electrophoresed on ABI PRISM™ 3730 DNA analyser with appropriate size standard by the Core Sequencing Facility (School of Medicine and Biomedical Sciences, University of Sheffield, Sheffield, UK). The data were analysed using genemapper software version 4.0 (Applied Biosystems).

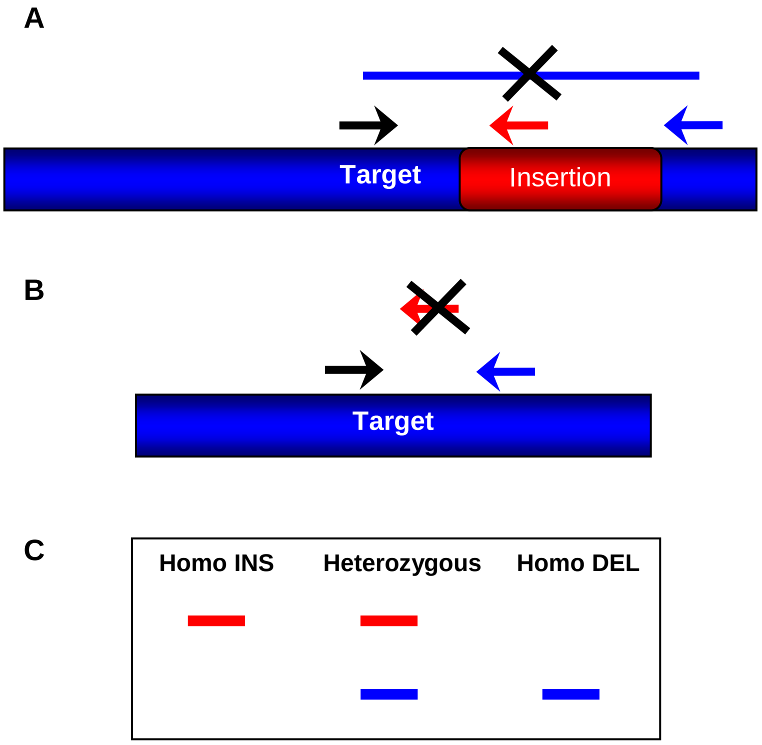


Figure 2.2 GAP PCR

A schematic presentation of GAP PCR. Arrows represent the primers used. Black is the common forward, Red is the insertion reverse and Blue is the deletion reverse A) The target sequence with an insertion, the reverse deletion will anneal, however, the PCR conditions will not allow the amplification of large amplicons. B) The target sequence with the deletion, the insertion reverse will not anneal C) The different genotypes as will appear on agarose gel.

2.2.7 Agarose gel electrophoresis

Prior to further analysis, PCR products were separated and identified using agarose gel electrophoresis. To prepare an agarose gel, the following procedure was used. Depending on the agarose percentage and the volume of gel required (usually 1.5% in 100 ml), an appropriate mass of Multi-Purpose Agarose (Bioline Ltd., London) and an appropriate volume of 1x TAE were mixed in a conical flask and heated in the microwave oven in short bursts. Once the agarose had fully dissolved and the solution had cooled down, 5 μl (10 mg/ml) ethidium bromide (Et.Br.) was added. Cooling the gel before Et.Br. addition reduces vaporisation of this mutagenic reagent. The solution was then poured into a gel casting tray (with combs in position) and left to solidify. When ready, the gel was placed in a Mini-Sub® Cell gel tank (Bio-Rad Laboratories Ltd., Hemel Hempstead) and overlaid with TAE (1x) as buffering system. 5 μl Et.Br. (10 mg/ml) was also added to the running buffer.

To determine the size of the PCR product, a standard size marker was electrophoresed alongside the PCR products. According to the expected size of the PCR product, two different standard size DNA markers were used, hyperladder I and IV supplied by Bioline-London (Appendix 4). 5 μl of the PCR products amplified with the Reddy Mix (contains a loading dye) were loaded directly into the wells. 5 μl of the long range PCR products were mixed with 1 μl DNA loading buffer (6x) (0.25% (w/v) bromophenol blue, 30% (v/v) glycerol) prior to loading. Gels were electrophoresed at 100 V for 35 min for conventional PCR products and at 120 V for 1 hour for long-range PCR products. DNA bands were visualised using a UV trans-illuminator U:Genius from Syngene (Cambridge).

2.2.8 Quantitative fluorescent PCR & microsatellite instability analysis

Quantitative fluorescent PCR (QF-PCR) is based on the PCR amplification of polymorphic microsatellites using fluorescently labelled primers. The products are visualised and quantified by capillary electrophoresis on an automated genetic analyser. The microsatellite instability (MSI) status of all the tumour and normal DNA samples analysed in this study was determined using a commercially available QFPCR kit: MSI Analysis System, Version 1.2 Kit (Promega). The kit includes fluorescently labelled primers for five quasimonomorphic mononucleotide microsatellite markers; BAT-25, BAT-26, NR-21, NR-24 and MONO-27 as recommended by the National Cancer Institute (NCI) revised criteria on MSI testing (Bacher *et al.* 2004, Umar *et al.* 2004). The MSI analysis was performed according to the manufacturer's protocol: 2 ng DNA was amplified in a 10 µl PCR mix containing; 5.85 µl nuclease free water, 1.0 µl Gold STAR 10x buffer, 1.0 µl 10x multiplex primer mix and 0.15 µl Amplitaq Gold DNA polymerase (5 U/µl) (Applied Biosystems).

The MSI cycling conditions were performed in a GeneAmp® PCR System 9700 thermocycler (Applied Biosystems) and consisted of the following steps; initial denaturation at 95^{o}C for 11 min, followed by another denaturation step at 96^{o}C for 1 min, then 10 cycles of denaturation (100% ramp) at 94^{o}C for 30 sec, annealing (29% ramp) at 58^{o}C for 30 sec and extension (23% ramp) at 70^{o}C for 1 min, followed by 20 more cycles of denaturation (100% ramp) at 90^{o}C for 30 sec, annealing (29% ramp) at 58^{o}C for 30 sec and extension (23% ramp) at 70^{o}C for 1 min, then a final incubation step at 60^{o}C for 30 min. The products were then analysed on ABI PRISM^{TM} 3730 DNA Analyser with appropriate size standard by the Core Sequencing Facility. The results were analysed using genemapper software version 4.0 (Applied Biosystems). The MSI markers were considered unstable if they exhibit a change in size in the tumour DNA compared to the paired normal DNA. Samples were classified to have high MSI (MSI-H) if they showed instability in two or more of the 5 mononucleotide markers, low MSI (MSI-L) if they showed instability in one of the markers, and MSS if none of the markers showed instability (Murphy *et al.* 2006)

2.2.9 TaqMan real time PCR

TaqMan assays involve the use of dual-labelled oligonucleotide probes complimentary to the target DNA sequence as illustrated in Figure 2.3 (Heid *et al.* 1996). The probes are fluorescently labelled with a reporter fluorophore at the 5' end and a 'quencher' molecule at the 3' end. The quencher molecule quenches the fluorescent signal emitted by the fluorophore. The quenching process requires close proximity between the quencher and the reporter molecules. During PCR, the probe will anneal to the target sequence. However, due to the 5' exonuclease activity of Taq polymerase, the probe, if perfectly annealed, will be hydrolysed resulting in the release of the reporter and quencher molecules and consequently the generation of a fluorescent signal. Subsequently, the signal is detected and quantified using a quantitative PCR machine (Figure 2.3). If the probe is not perfectly matched to the target area, it will not be hydrolysed and the fluorescent signal will be minimal as the 3' quencher stays in close proximity to the 5' fluorophore.

2.2.9.1 TaqMan genotyping assays

TaqMan assays were used for the genotyping of a 6bp insertion-deletion (ins/del) polymorphism (rs3834129) in the promoter region of *CASP8*. Probes specific for each allele were used and each probe was labelled with a different reporter fluorophore (FAM^{TM} for insertion and VIC® for deletion). Thus, depending on the attached probe, the signal differs according to the genotype at that position. PCR reactions were performed in 384-well PCR plates with 10 ng genomic DNA in each well. Standard 5 µl PCR reactions were used. Each reaction contained 2.5 µl of 2x TaqMan Genotyping mastermix (ultrapure hot start AmpliTaq Gold DNA polymerase, deoxynucleotide triphosphates with UTP, AmpErase uracyl-*N*-glycosylase (UNG), passive reference dye, and optimised buffer), 0.0625 µl of primers and TaqMan probe mix (Applied Biosystems) and 2.4375 µl deionised water. The plates were sealed using BioRad PCR microseals (Bio-Rad Laboratories Ltd.).

After centrifugation of the PCR plates for 1 min at 2000 RPM, TaqMan PCR reactions were performed under the following conditions: initial incubation at 50^{o}C for 2 min, initial denaturation at 92^{o}C for 10 min, followed by 40 cycles of denaturation at 92^{o}C for 15 sec and combined annealing/extension at 60^{o}C for 1 min. All TaqMan assays were carried out in a Dyad Peltier Thermal Cycler (MJ Research, currently supplied by Bio-Rad Laboratories Ltd.). TaqMan PCR products were analysed using an ABI HT7900 real time PCR machine (Applied Biosystems). The data were analysed using the allelic discrimination module of the Sequence Detector Software version 2.2.1 (Applied Biosystems).

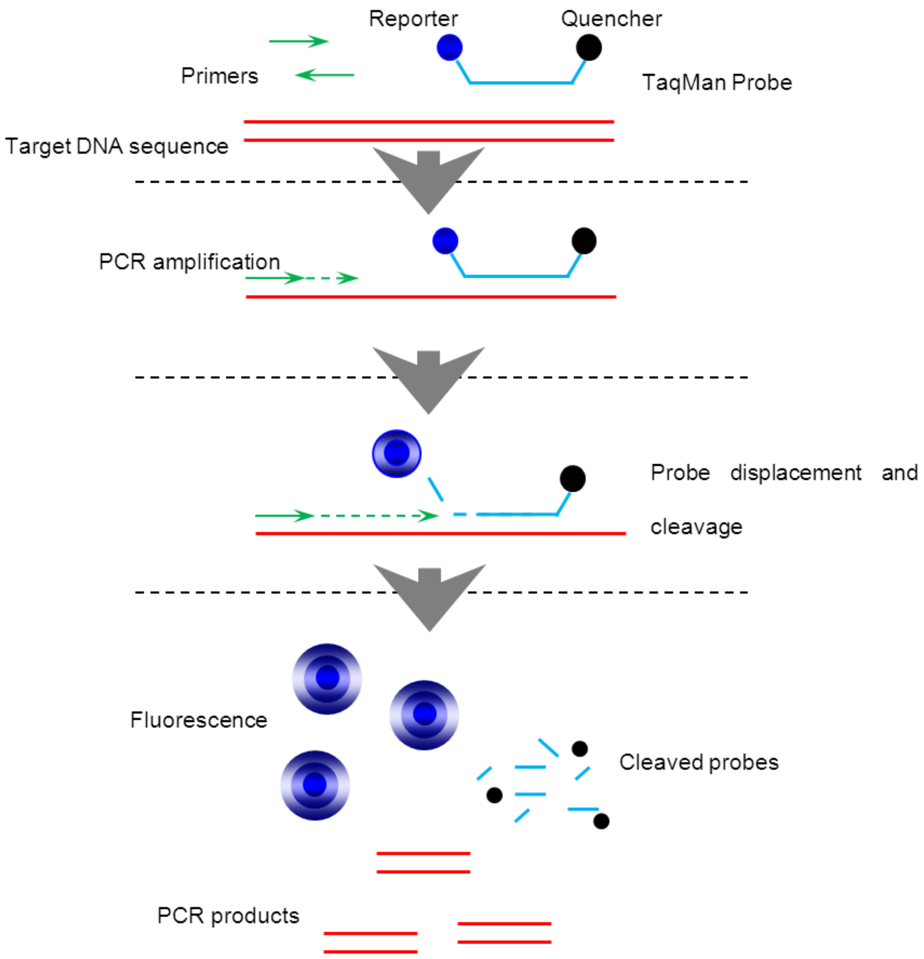


Figure 2.3 TaqMan assay principle

A schematic presentation of the TaqMan assay principle. For simplicity, the figure is showing PCR amplification in the forward direction only.

2.2.9.1.1 TaqMan genotyping quality control

Several measures were taken to ensure good quality genotyping results. For example, up to 10% of the DNA samples were duplicated on separate 384-well PCR plates. These samples were thus genotyped twice to check the reproducibility of the assay. Duplicate samples were considered informative when successfully called in both assays and a duplicate concordance rate of ~98% represented high reproducibility. Additionally, no-template controls (NTC) were included on each plate to check for DNA contamination.

2.2.9.2 TaqMan copy number assays

Real Time PCR TaqMan copy number assays were used to validate the accuracy of the aCGH data (Section 2.2.15). Three of the common focal aberrations (*PARK2*, *KCNMA1* and *NFKBIA*) identified in this project were chosen for confirmation by TaqMan copy number assays. Pre-designed assays were chosen from the ABI website (http://www.appliedbiosystems.com) (Table 2.5). Pre-designed TaqMan Copy number reference assays (*TERT* and *RPPH1*) were also used for the relative quantification of the copy numbers of the targets. The assays contain forward and reverse primers and a dye-labelled TaqMan probe (FAM^{TM} for the target assays and VIC® for the reference assays). These reference assays were chosen to anneal to a region in the sample known to have a diploid copy number, based on the aCGH results. The target copy number assays were run simultaneously with the copy number reference assay in a duplex real time PCR. The number of copies of the target sequence is determined by relative quantification (RQ) based on a comparative threshold cycle (C_T) method ($\Delta\Delta C_T$). Threshold cycle is the fractional cycle number at which the emitted fluorescence level passes the detection threshold (Figure 2.4). The C_T difference (ΔC_T) between the target and the reference assays is calculated and compared to the ΔC_T value of control DNA samples known to have a chromosomally stable diploid genome.

Table 2.5 TaqMan copy number assays

Gene	Assay Number	Cytoband	Genomic location	Genomic Build
PARK2	Hs03585570-cn	6q26	chr6:162636459	37.1
KCNMA1	Hs03760015-cn	10q22.3	chr10:79111001	37.1
NFKBIA	Hs00462687-cn	14q13.2	chr14:35870871	37.1

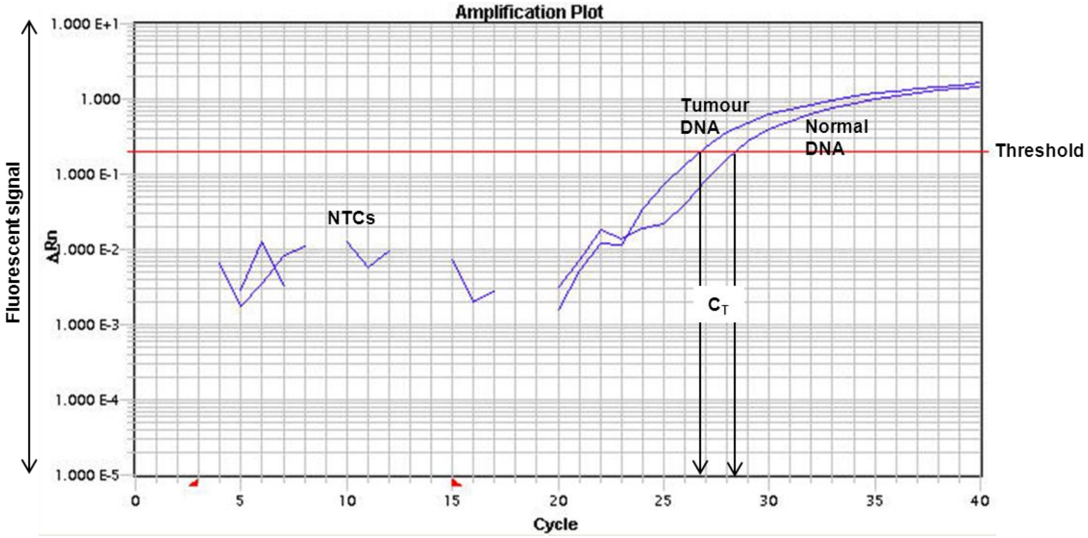


Figure 2.4 TaqMan copy number assay amplification plot

A TaqMan copy number assay amplification plot showing tumour and normal DNA amplification curves. The C_T of the tumour DNA is smaller in comparison to the C_T of the normal DNA indicating an increased copy number in the Tumour DNA. This sample had amplification at this genomic location.

PCR reactions were performed in 384-well PCR plates with 2 µl of 5 ng/µl DNA samples. Standard 8 µl PCR reactions were used. Each reaction contained 5 µl of 2x TaqMan Genotyping mastermix, 0.5 µl of 20x TaqMan copy number target assay, 0.5 µl of the 20x TaqMan copy number reference assay and 2 µl nuclease free water. NTC were also included in all of the experiments to check for contamination. The plates were sealed using BioRad PCR microseals (Bio-Rad Laboratories Ltd.). After centrifugation of the PCR plates for 1 min at 2000 RPM, they were loaded in an ABI HT7900 real time PCR machine (Applied Biosystems) and reactions performed using the following conditions; initial denaturation at 95°C for 10 min, followed by 40 cycles of denaturation at 95°C for 15 sec and combined annealing/extension at 60°C for 1 min. The results were then initially analysed using the copy number module of the Sequence Detector Software version 2.2.1 (Applied Biosystems). Copy numbers for the samples were then predicted using the Copy Caller software v1.0 (Applied Biosystems).

2.2.10 DNA sequencing

During this project, bidirectional automated dye terminator DNA sequencing was used to identify any sequence variants in specific target regions. This DNA sequencing method is based on dideoxynucleotide (ddNTPs) sequencing chemistry developed by Frederick Sanger in the 1970s (Sanger *et al.* 1977). In the current evolved form of Sanger sequencing, 4 different fluorescently labelled ddNTPs (one for each of the 4 bases adenine, thymine, cytosine, and guanine) are mixed with non-labelled dNTPs (A, T, C and G) in a PCR reaction. The ddNTPs incorporation during the PCR extension step results in termination of the reaction as a result of the absence of a 3'-OH group. This will happen at all positions in the target DNA sequence and result in a series of fluorescently labelled DNA fragments of different lengths. Capillary electrophoresis can then be used to separate the DNA fragments according to their length and the fluorescent signal will be captured at all positions. The last two steps are usually performed in an automated DNA analyser.

DNA sequencing was applied on PCR products for the detection of sequence variants. After being amplified and assessed (using agarose gel electrophoresis), PCR products were enzymatically purified in preparation for sequencing. This step was performed to remove the residual primers and to degrade the remaining dNTPs; both of which might interfere with the quality of the sequences obtained. Enzymatic purification was carried out using the following protocol: 6 µl PCR product, 4 µl of ExoSAP-IT (GE Healthcare UK Ltd., Little Chalfont) (1:4 dilution v/v) and 4 µl deionised H_2O were mixed in a 0.5 µl PCR tube. The reaction was then incubated at 37ºC for 30 minutes and denatured at 80ºC for 15 minutes. ExoSAP-IT is a mixture of two enzymes: the first one is Exonuclease I (Exo I) which removes any single-stranded DNA and primers. The second enzyme is Shrimp Alkaline Phosphatase (SAP) which degrades residual dNTPs (Werle *et al.* 1994; www.usbweb.com).

All sequencing reactions were carried out by the Core Sequencing Facility using an ABI PRISM^{TM} BigDye Terminator v3.1 Sequencing standard kit and were performed on an ABI PRISM^{TM} 3730 DNA analyser (Applied Biosystems).

2.2.11 Sequence analysis

2.2.11.1 Manual analysis

All sequences were visually checked for double peaks, which might indicate a heterozygous position. When necessary, the sequences were checked against the published gene sequences and control sequencing results.

2.2.11.2 Staden sequence checking

The sequences were also analysed with reference to control sequence traces using the Staden sequence analysis package (Bonfield *et al.* 1998). A blank Staden database was set up for *CASP8*, *APC*, *TP53*, *KRAS*, *BRAF* and *PIK3CA* analysis. The target sequence was exported and saved from the Ensembl database as a text file (www.ensembl.org, accessed September, 2008 – August, 2010). Next, pregap4 software (v1.4b1) from the Staden package was used to add the target sequence and a new database template was created. Subsequently, GAP4 software (v4.8b1) was used to edit the target sequence. All the exons were highlighted using the contig editor. For sequence analysis, the ab1 files of the generated sequences were placed in the Staden databases. The files for the samples and the controls were added using pregap 4 and the control was set as the reference sequence. GAP4 was then used to analyse the imported sequences. GAP4 compares the fluorescent traces of the target sequence against those of the control sequence. It then subtracts the pairs of wild-type and mutant traces to produce new traces that represent the differences; those that are significant are highlighted (Bonfield *et al.* 1998). Each amplicon was processed and analysed separately.

2.2.12 Methylation Specific Multiplex Ligation Probe Amplification

Multiplex Ligation Probe Amplification (MLPA) is a multiplex fluorescent PCR method that is used to detect copy number variation (Schouten *et al.* 2002) (Figure 2.5). The technology involves the use of oligonucleotide probes of variable, but unique, lengths complimentary to various DNA targets. Other than the target specific sequence, the MLPA probes will also have identical 5' or 3' end sequences that are then amplified by a single fluorescent primer pair. The products are analysed by capillary electrophoresis and different product sizes will reflect the different targets. In methylation specific MLPA (MS-MLPA), the addition of a digestion step using methylation sensitive endonucleases result in the ability to semi-quantify the methylation level in DNA samples (Nygren *et al.* 2005) (Figure 2.5). MS-MLPA has several advantages over other methylation analysis technologies; it is not based on bisulphite conversion, it requires a small amount of DNA, and it allows the multiplexing of several markers in a single reaction (Jeuken *et al.* 2007). The MS-MLPA kit ME042-A1 was used in this project to determine the methylation status of 8 markers (*CACNA1G*, *CDKN2A*, *CRABP1*, *IGF2*, *MLH1*, *NEUROG1*, *RUNX3* and *SOCS1*) in paired normal and tumour DNA samples. All the steps were carried according to the manufacturer's instructions (MRC-Holland, Amsterdam). It is noteworthy here that copy number analysis using the MS-MLPA kit was not performed, since the reference probes targeted unstable genomic areas in CRC.

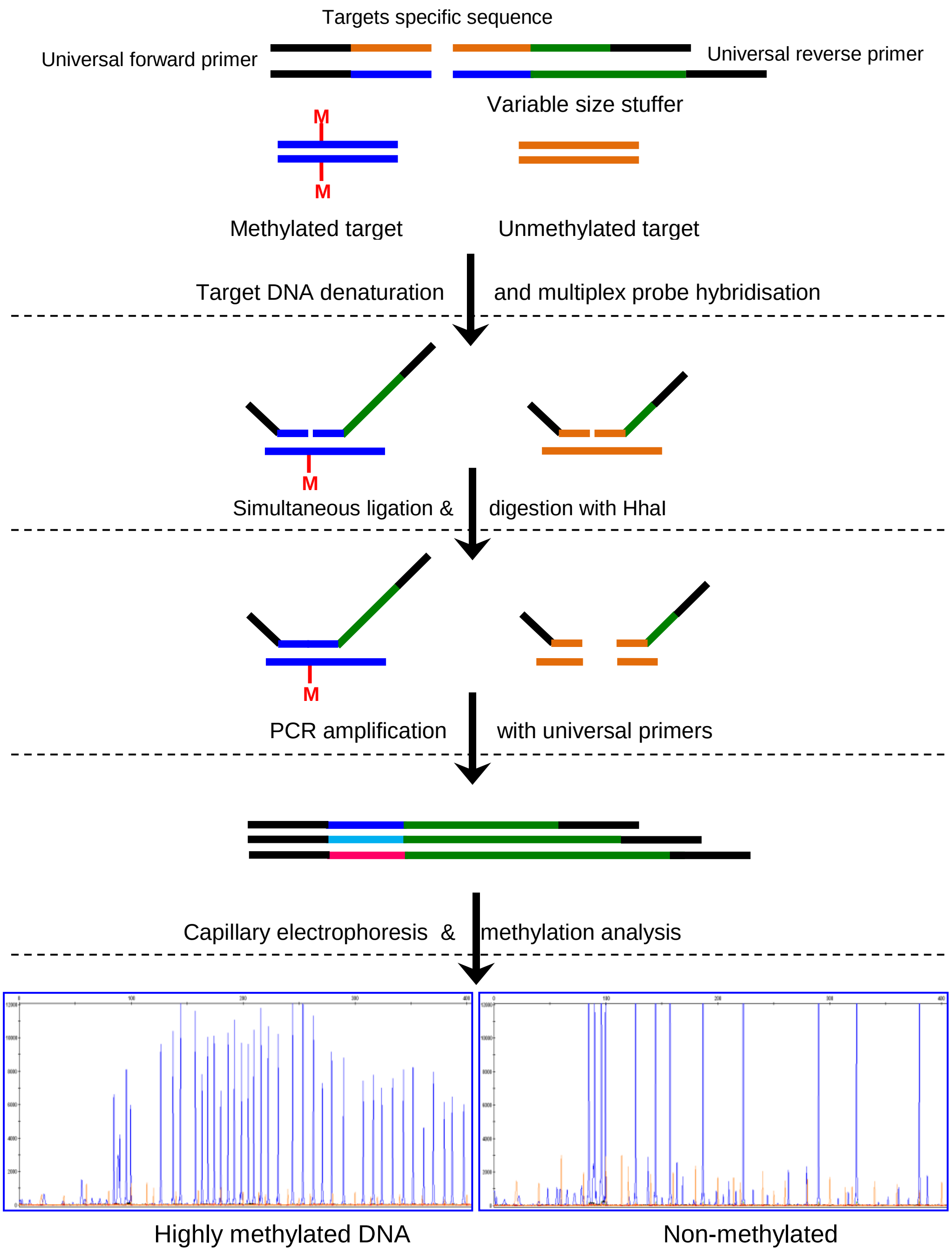


Figure 2.5 MS-MLPA technology

MS-MLPA probes contain 2 oligonucleotides; one short synthetic and one long M13-derived. Up to 50 MS-MLPA probes can be used in a single reaction. Both oligonucleotides contain universal primer sites and target specific sequences. Each M13 derived oligonucleotide also contains a “stuffer” sequence with a probe specific length. Moreover, MS-MLPA probes also contain an *HhaI* recognition site (only methylation specific probes). When attached to a methylated DNA target, methylation sensitive *HhaI* enzyme will fail to digest the probe, a fluorescent signal will be produced, and peaks will be observed through the genemapper analysis. When attached to unmethylated DNA target, *HhaI* will digest the probes and no signal will be detected. *Adapted from (Nygren *et al.* 2005).

2.2.12.1 DNA denaturation and MS-MLPA probe hybridisation

DNA samples (50 ng) were diluted in 5 µl of molecular biology grade 1xTE buffer (Promega) in 0.2 ml PCR tubes. The diluted DNA samples were then denatured by incubation in the GeneAmp® PCR System 9200 thermocycler (Applied Biosystems) at 98^{o}C for 10 min followed by cooling down to 25^{o}C. Following that, 3.0 µl of SALSA probe/MLPA buffer (1:1) mix were added to the samples, which were then mixed carefully by pipetting up and down and incubated at 95^{o}C for 1 min followed by 60^{o}C for 16 hrs.

2.2.12.2 Ligation and digestion

Before the end of the 16 hr incubation, the following mixes (per sample) were prepared and kept on ice; Ligase buffer A mix: 3 µl of ligase buffer A with 10 µl of nuclease free water, ligase-65 mix: 1.5 µl of ligase-65 buffer B, 0.25 µl of ligase-65 enzyme and 8.25 µl of nuclease free water, ligase-digestion mix: 1.5 µl of ligase-65 buffer B, 0.25 µl of ligase-65 enzyme, 0.5 µl *HhaI* enzyme (10 U/ µl) and 8.25 µl of nuclease free water. For each sample, a new 0.2 ml PCR tube was labelled and prefixed with D (referring to digestion mix). After the end of the 16 hr hybridisation, 13 µl of ligase buffer A mix were added to each sample tube and mixed by pipetting. After that, 10 µl of the sample/ligase buffer A mix were transferred to the D prefixed labelled PCR tubes. Both tubes were then incubated at 49^{o}C for at least 1 min (in the thermocycler). While at 49^{o}C, 10 µl ligase-65 mix were added to the first tube of each sample (ligation reaction), then 10 µl of the ligase-digestion mix were added to the D prefixed labelled tubes of each sample (ligation digestion reaction). Both tubes were then incubated at 49^{o}C for 30 min followed by 5 min at 98^{o}C.

2.2.12.3 Amplification and capillary electrophoresis

For each sample, 2 PCR tubes (0.2 ml) were labelled (with and without the D- prefix). The following mixes were prepared and kept on ice; SALSA PCR buffer mix: 2 µl of SALSA PCR buffer with 13 µl nuclease free water, polymerase mix: 1 µl SALSA PCR primers, 1 µl SALSA enzyme dilution buffer, 0.25 µl SALSA polymerase and 2.75 µl nuclease free water. In the new labelled 0.2 ml PCR tubes, 5 µl of MLPA ligation or 5 µl of ligation digestion reaction were mixed with 15 µl of SALSA PCR buffer mix. The tubes were placed on ice and 5 µl polymerase mix was added to each tube. The tubes were then placed in a preheated thermocycler (72^{o}C) and amplified using the following PCR conditions; 35 cycles of denaturation at 95^{o}C for 30 sec, annealing at 60^{o}C for 30 sec and extension at 72^{o}C for 60 sec followed by 20 min incubation at 72^{o}C. The PCR products were then electrophoresed on an ABI PRISM^{TM} 3730 DNA Analyser with appropriate size standards by the Core Sequencing Facility.

2.2.12.4 MS-MLPA results analysis

The initial analysis of the MS-MLPA products was performed using genemapper software version 4.0 (Applied Biosystems). The raw data was then exported as a text file and the rest of the analysis described in Sections 2.2.12.4.1 - 2.2.12.4.3 was performed in Microsoft Excel spreadsheets templates and using an in-house algorithm designed by Dan Connely (Bioinformatician in our research group).

2.2.12.4.1 Quality control

Two types of quality control (QC) fragments are included in the MS-MLPA kit, the “Q” and the “D” fragments. The Q-fragments (64, 70, 76 and 82bp) are generated by the Q-probes in a non-ligation dependent manner. The Q-peaks areas should be smaller than half the area of D-peaks (below), if they were not, then the sample ligation failed or the DNA concentration was very low. The D-fragments (88, 92 and 96bp) are generated by the D-probes in a ligation dependent manner. The D-peaks areas should be comparable to other MLPA reference probes areas (≥40%) which mean that the ligation was successful, there was a sufficient amount of DNA and that they were properly denatured. If the 88 and the 96 D-fragments averaged areas were <40% of the 92 D-fragment area then denaturation was not enough. If the area of the 92 D-fragment was <40% of the 88 and 96 D-fragments averaged areas, it indicates that hybridisation was incomplete. This can result from short hybridisation time, low hybridisation temperature, insufficient amounts of MLPA probe mix and/or MLPA buffer, the use of >5ul of DNA, or the thermal cycler lid temperature was less than 100^{o}C. The QC tests are summarised in Table 2.6. Finally, a digestion control probe is included in the probe mix (Section 2.2.12.1) to test the efficiency of the digestion step. This probe should not give a signal upon digestion.

Table 2.6 MS-MLPA "Q" and "D" fragments QC tests

	Test	Description	Pass*	Fail*
A	**DNA Concentration/Ligation**	Q-peaks against D-peaks	≤50%	>50%
B	**Denaturation**	88D and 96D against 92D	≥40%	<40%
C	**Hybridisation**	92D against 88D and 96D	≥40%	<40%
D	**DNA Quantity/Ligation/Denaturation**	D-peaks against reference peaks	≥40%	<40%

* The % reflect the ratio between the peaks areas

2.2.12.4.2 Intra-sample data normalisation

For the analysis of the MS-MLPA results, relative peak areas are used instead of absolute peak areas (generated by the ABI PRISM^{TM} 3730 DNA Analyser). In order to calculate the relative peak areas, the data was normalised in an intra-sample fashion. This involved dividing the peak area generated from each marker probe by the peak area from each of the 10 reference probes in the same sample, thus creating 10 ratios for each of the marker probes. The median of the 10 marker probes ratios is considered the normalisation constant of the probe and is used for the methylation and copy number analyses. This normalisation was performed for the undigested as well as the digested samples.

2.2.12.4.3 Methylation analysis

The methylation status of the marker probe regions is calculated by dividing the normalisation constant of the probes from the digested sample by the normalisation constant of the probes from the corresponding undigested sample and multiplying the result by 100 to obtain the methylation percentage. In order to identify aberrant methylation, the tumour DNA methylation percentages were compared to their matching normal DNA controls. A marker is considered to be methylated in the tumour DNA if the averaged methylation levels of all of its probes were greater by ≥10% than the value for the normal DNA. However, at least 2 probes should indicate such an increase in methylation levels. This level of increase was previously shown to correlate with a marked decrease in mRNA expression (Cheng *et al.* 2008). Finally, a sample was classified as CIMP negative (CIMP-N) if none of the markers were methylated, CIMP low (CIMP-L) if 1-4 of the markers were methylated and CIMP high (CIMP-H) if 5 or more were methylated (Ogino *et al.* 2007).

2.2.12.4.4 Sex determination and *BRAF* V600E mutation

In addition to copy number and methylation analysis, the MS-MLPA kit ME042-A1 provides X and Y chromosome specific probes which can be used to determine the gender of the samples. Moreover, a mutation-specific probe that binds only in the presence of the *BRAF* mutation V600E is available.

2.2.13 DNA bisulphite treatment

DNA bisulphite treatment results in the conversion of unmethylated cytosine residues to uracil with no effect on methylated cytosine (5-methylcytosine) (Ehrich *et al.* 2007). Bisulphite treatment of DNA is an essential step in many methylation studies as it is used prior to techniques including methylation specific PCR, real time PCR, sequencing and pyrosequencing.

In this project, bisulphite treated DNA was used in pyrosequencing reactions to validate methylation positive and negative controls for MS-MLPA tests. The CpGenome DNA modification kit from Chemicon (Millipore, Watford) was used to perform DNA bisulphite treatment. All the reactions were performed according to the manufacturer’s protocol. Briefly, 500 ng DNA was diluted to 10 µl in nuclease free water and then mixed with 97 µl of DNA modification reagent 4 (R4) mix (alkaline solution). The DNA/R4 mix was then incubated at 50^{o}C for 10 min. This results in the denaturation of the DNA (mild heat at an alkaline pH). 550 µl of DNA modification reagent 1 (R1) is then added to the samples and the mixture is incubated at 50^{o}C for 16 hr in the dark. R1 contains sodium bisulphite, which results in the conversion of unmethylated cytosine into uracil sulfonate. After 16 hr, 5 µl of DNA modification reagent 3 (R3) were added to the samples, followed by 750 µl of DNA modification reagent 2 (R2). R2/3 mix results in the binding of the DNA to micro-particulate carriers.

Samples were then incubated at room temperature for 10 min, centrifuged at 8000 *x*g for 1 min and the supernatant was carefully discarded. The pellet was then washed 3 times by adding 70% ethanol, vortexing and spinning at 8000 *x*g for 1 min. After the last wash, 50 µl of 20 mM NaOH/90% ethanol was added, and the tubes were incubated at room temperature for 5 min and then washed twice with 90% ethanol. This results in DNA desalting and the complete conversion of the intermediate uracil sulfonate into uracil (*via* alkaline desulfonation). After removing the supernatant, the samples were air dried for 20 min. The pellet was then dissolved in 20 µl 1x TE buffer (promega), vortexed and incubated at 60^{o}C for 15 min. The samples were then centrifuged at 13000 *x*g for 3 min and the solution containing the bisulphite treated DNA transferred to new tubes.

2.2.14 Pyrosequencing

Pyrosequencing is a “sequencing by synthesis” method that is dependent on detecting pyrophosphate release during nucleotide incorporation (Brakensiek *et al.* 2007). Pyrosequencing is only used for sequencing known and short DNA sequences, to screen for known variations (mutations or SNPs), or to quantify methylation levels in bisulphite treated DNA. In this project, pyrosequencing was used to validate the methylation status of DNA from HCT116 cells, in addition to methylation positive and negative controls used to test the MS-MLPA technology.

2.2.14.1 Pyrosequencing primer design

Pyrosequencing primers for 2 of the methylation markers included in the MS-MLPA kit (*NEUROG1* and *CRABP1*) were designed using the software PyroMark assay design v2.0 (Qiagen). The assay type was selected as methylation analysis (CpG). DNA sequences of both markers were imported into the software and the target regions were selected. PyroMark assay design software provided lists of 3 primers for each region; 2 PCR primers (one biotin labelled) and a sequencing primer. The genomic areas underneath the primers were then checked for SNPs using the diagnostic SNP check online tool provided by NGRL (http://ngrl.man.ac.uk/SNPCheck/SNPCheck.html, accessed April, 2010).

2.2.14.2 Pyrosequencing PCR

Pyrosequencing PCR was performed using hot start Taq polymerase (Qiagen). Standard 25 µl PCR reactions were used; each reaction contained 1-2 µl of the bisulphite treated DNA, 2.5 µl of 10x PCR buffer, 1.5 µl $MgCl_2$ (25mM), 0.5 µl dNTPs, 0.2 µl HotStart Taq and up to 25 µl deionised water. PCR cycling conditions were; initial denaturation step at 95ºC for 15 min, followed by 50 cycles of denaturation at 95ºC for 20 sec, annealing at nºC for 20 sec and extension at 72ºC for 20 sec, and a final extension step at 72 ºC for 5 min. The annealing temperatures varied according to the T_M of the primers (Appendix 1, Table 8). All PCR reactions were carried out using GeneAmp® PCR System 9200 thermocycler (Applied Biosystems). Prior to pyrosequencing, all the PCR products were analysed using agarose gel electrophoresis (Section 2.2.7)

2.2.14.3 Pyrosequencing reaction

The PCR products were taken to the Sheffield Children's hospital where the pyrosequencing procedures were performed. The pyro-CpG software was used to create the specific assay files for the reactions included in the run. Capturing master mixes were firstly prepared by mixing 30 µl deionised water, 38 µl binding buffer and

2 µl sepharose beads for each sample. The capturing mix (70 µl) was then added to the corresponding wells and 10 µl of the PCR products were then added to the wells. The plate was then sealed and transferred to a shaker and incubated for 10 min at 1400 RPM. During this step, the biotin labelled PCR products are attached to the sepharose beads. During the 10 min incubation step, pyrosequencing master mix was prepared by mixing 0.36 µl sequencing primers (10 pmol/ µl) with 11.64 µl of the annealing buffer. The pyrosequencing master mix was then transferred to the corresponding wells of a pyrosequencing plate and the plate was placed in the designated position of the pyrosequencing working station (Figure 2.6). The pyrofilter was hydrated in deionised water and checked for blocked filters by observing the water levels in the wells after switching on the vacuum pump. This pyrofilter is an aspiration device that is used to capture the biotin labelled PCR products attached to the sepharose beads. After 10 min shaking, the PCR plate was unsealed and transferred to the designated position on the working station (Figure 2.6).

The vacuum pump was switched on and the pyrofilter was used to aspirate all the PCR products from the wells. Any biotin labelled PCR products attached to sepharose beads will be captured by the filters tips and all the other PCR reagents will pass through. The captured biotin labelled PCR products were denatured and washed by immersing the pyrofilter tips (containing captured PCR products) into the 3 successive stations (~5 sec each) of ethanol, denaturation solution and washing buffer. At the end of this step, washed single stranded biotin labelled DNA sequences are attached to the pyrofilter tips ready to be released onto the pyrosequencing plate. The pyrofilter was positioned on top of the pyrosequencing plate, the pump was switched off and the pyrofilter was then lowered into the pyrosequencing plate wells. A gentle shake was the used to release the biotin labelled single stranded DNA. The pyrofilter was transferred to the water station and the sequencing plate was incubated at 80^{o}C on a hot plate for a few minutes.

While the plate is incubating, the pyrogold reagents were prepared. The required amounts of the substrate, enzyme and nucleotides (A, C, G and T) were added into the respective tips. While adding the reagents to the tips, air bubbles were strictly eliminated. The tips were checked for blockage by gently pressing the sides after covering the top of the tips with gloved fingers. A circular tiny drop should appear. If not, then the tip is blocked or there is an air bubble. After checking, tips were placed in the tip-holder in the correct order (Figure 2.7) and it was transferred to the pyrosequencer (PyroMark^{TM} MD) to test the dispensation. The tip-holder was placed in the specific chamber and dispensation was tested using a sealed plate underneath the tips. The sealed plate should have 6 drops after dispensation coming from the substrate, enzyme and the 4 nucleotides tips. After the dispensation test, the pyrosequencer was ready to run and the pyrosequencing plate was transferred into position and the reaction started. When the run was completed, the pyrosequencing data were analysed using the pyro-CpG software (Qiagen).

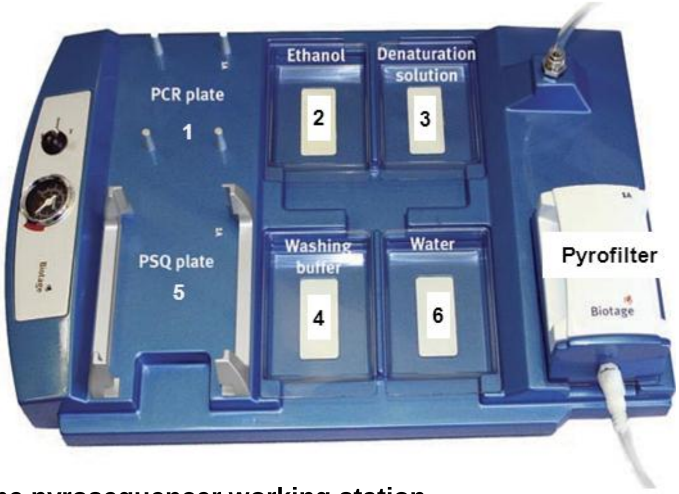


Figure 2.6 The pyrosequencer working station

The pyrofilter will be used in station 1 to aspirate the solution from the PCR plate wells. Biotin labelled PCR products attached to sepharose beads will be captured by the pyrofilter tips. The PCR products will then be denatured and washed (stations 2-4). Biotin labelled ssDNA will then be released into the pyrosequencing (PSQ) plate in station 5 and then the pyrofilter will be washed in water (station 6) (Figure adapted from http://www.pyrosequencing.com/).

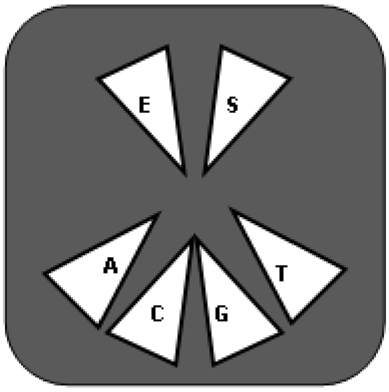


Figure 2.7 The pyrosequencer tip holder

The pyrosequencer tip holder showing the locations of the enzyme (E), substrate (S) and the four nucleotides.

2.2.15 Array Comparative Genome Hybridisation

Array comparative genome hybridisation (aCGH) is an array-based technology that is used to compare genomic copy number variation between two DNA samples at high resolution. It is widely used in cancer genetics to detect copy number changes acquired by tumour cells by comparing normal and tumour DNA (Pinkel and Albertson 2005). Figure 2.8 summarises the aCGH technology. Briefly, tumour DNA is usually isolated from the cancer cells and control DNA is ideally isolated from normal cells from the same patient if available (e.g. peripheral blood). If not available, commercial sex-matched pooled genomic DNA can be used. Equal amounts of the tumour and normal DNA are first labelled with different fluorescent dyes, mixed and then hybridised to an array slide. The microarray slides contain oligonucleotide probes referring to human genomic sequences. After hybridisation and washing, the array slides are then scanned and the fluorescent intensities are measured from each probe. The scanned images are then interpreted and quantified using feature extraction (FE) software. The output from the FE software is then analysed using copy number analysis software where the ratios of the different dyes will be interpreted for each genomic region. All the aCGH steps were carried out according to Agilent Oligonucleotide Array-Based CGH for Genomic DNA Analysis version 6.0 (Agilent Technologies UK Ltd., Wokingham).

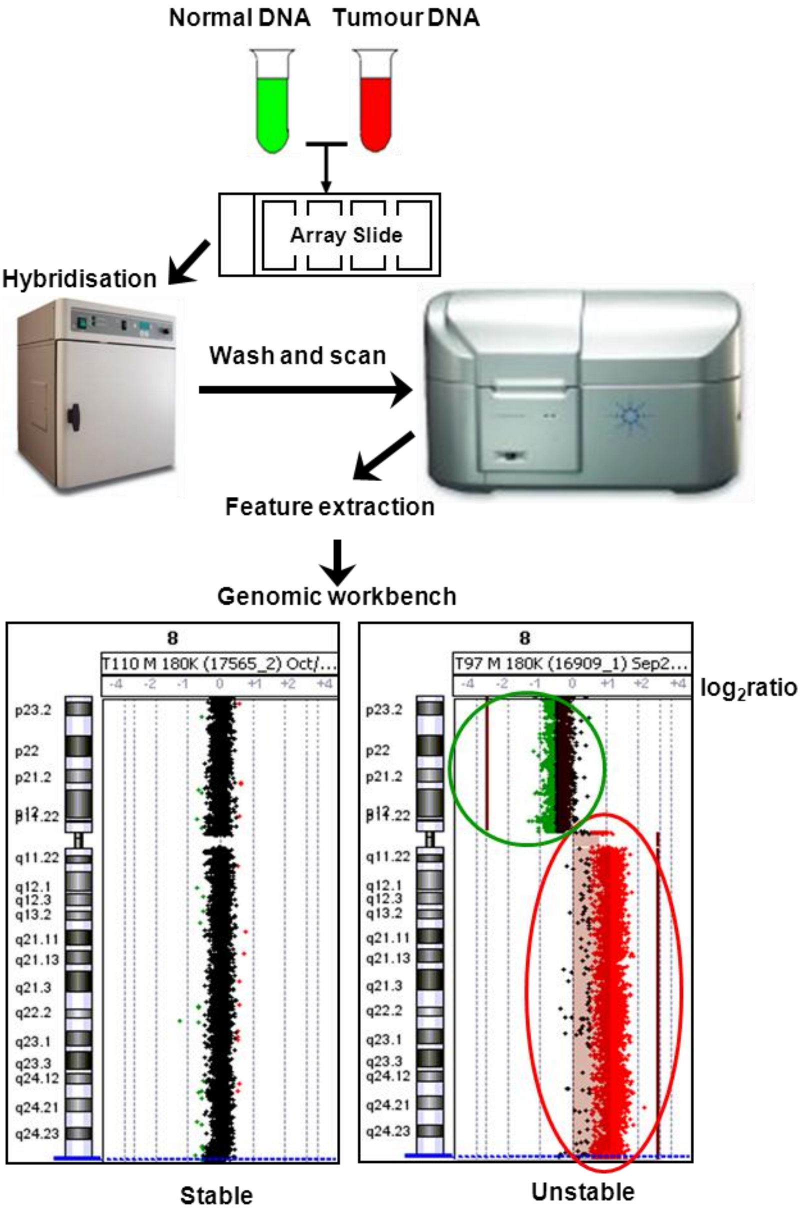


Figure 2.8 Schematic representation of the aCGH technology

Equal amounts of normal and tumour DNA are differentially labelled and hybridised to an array slide. After washing, scanning and feature extraction, the $\log_2$ratios of the normal and the tumour DNA signals are calculated. The dots underneath the $\log_2$ratio scale represent probes. Positive and negative $\log_2$ratios reflect amplifications (red probes) and deletions (green probes) respectively. Probes coloured in black represent stable genomic regions detected by $\log_2$ratios around zero.

2.2.15.1 Direct method

The direct method of aCGH was used for samples for which a minimum amount of 0.5 µg DNA was available. In this method, DNA samples were fragmented by restriction digestion using *AluI* and *RsaI*. The DNA samples (0.5 µg in 21 µl) were mixed with 5.8 µl digestion master mix which consisted of; 2 µl nuclease free water, 2.6 µl 10x buffer C (supplied with *RsaI*), 0.2 µl acetylated bovine serum albumin (BSA) (10µg/ µl), 0.5 µl *AluI* (10U/ µl) and 0.5 µl *RsaI* (10 U/µl). The mix was then incubated at 37^{o}C for 2 hr followed by 20 min at 65^{o}C. In order to evaluate the completeness of digestion, the samples were then analysed on 1.5% agarose gel.

2.2.15.2 Whole genome amplification

When less than 0.5 µg of DNA was available, WGA was used to produce a representative amplification of genomic DNA (Brueck *et al.* 2007). In this research project, WGA was performed using a commercially available kit GenomePlex (Sigma Aldrich). The first step was DNA fragmentation; 10 µl of the DNA samples (50 ng) were mixed with 1 µl 10x fragmentation buffer and incubated at 95^{o}C for exactly 4 min. Then, the samples were immediately cooled on ice. The second step was library preparation; 2 µl of 1x library preparation buffer and 1 µl of library stabilisation solution were added to the fragmented DNA samples and incubated at 95^{o}C for 2 min. The samples were then cooled on ice and centrifuged briefly. Next, 1 µl of the library preparation enzyme was added to each sample and mixed thoroughly by pipetting. The samples were centrifuged briefly and incubated at 16^{o}C for 20 min, 24^{o}C for 20 min, 37^{o}C for 20 min and 75^{o}C for 5 min in GeneAmp® PCR System 9200 thermocycler (Applied Biosystems). The last step of the WGA protocol was DNA amplification; 60 µl of amplification master mix was prepared for each sample. Each reaction contained 7.5 µl 10x amplification master mix, 47.5 µl nuclease free water and 5 µl WGA DNA polymerase. The amplification master mix was then added to the 15 µl from the library preparation step, mixed thoroughly and amplified with the following cycling conditions; initial denaturation at 95ºC for 3 min, followed by 14 cycles of denaturation at 94ºC for 15 sec, annealing/extension at 65ºC for 5 min.

2.2.15.2.1 Purification of WGA PCR products

Following WGA, the PCR products were purified using the commercially available kit GenElute PCR clean-up (Sigma Aldrich); the 75 µl WGA products were mixed with 375 µl binding solution and transferred to the miniprep binding column in a collection tube. The miniprep columns with the samples were then centrifuged at 13000 *x*g for 1 min. The eluate was discarded, 0.5 ml of the wash buffer was added to the miniprep columns (supplied in the kit) and they were centrifuged at 13000 *x*g for 1 min. The eluate was discarded again and the miniprep columns were centrifuged at 13000 *x*g for 2 min, to remove any residual wash buffer. Next, the miniprep columns were transferred to a fresh collection tube and 50 µl of elution buffer was added to the centre of each column. The columns were incubated at room temperature for 1 min and finally centrifuged at 13000 *x*g for 1 min to recover the purified WGA products in the eluate. The purified products were quantified using Nanodrop as described in Section 2.2.2. The WGA yield (µg) was calculated by multiplying DNA concentration (ng/ µl) by the sample volume (50 µl) and dividing by 1000.

2.2.15.3 Fluorescent labelling of DNA

After the restriction digestion or WGA, the normal and tumour DNA samples were fluorescently labelled with cyanine 3-dUTP and cyanine 5-dUTP respectively; Firstly, DNA samples were mixed with 5 ng of random primers and incubated at 95^{o}C for 3 min followed by 5 min on ice. Subsequently, the samples were centrifuged at 6000 *x*g for 1 min and the labelling master mix was prepared; 2.0 µl nuclease free water (only for restriction digestion products), 10 µl 5x buffer, 5.0 µl 10x dNTP, 3.0 µl cyanine 3-dUTP or cyanine 5-dUTP and 1.0 µl exo-klenow fragment. The labelling master mix was then added to 19 µl restriction digest product from Section 2.2.15.1 or 21 µl cleaned WGA product from Section 2.2.15.2.1 and then mixed thoroughly by pipetting. The cyanine 3-dUTP mix was used for the normal DNA and the cyanine 5-dUTP was used for the tumour DNA. The samples were then incubated at 37^{o}C for 2 hr, followed by 10 min at 65^{o}C for 10 min.

2.2.15.4 Clean-up of labelled genomic DNA

The labelled genomic DNA was purified using microcon YM-30 filters (Fisher Scientific UK Ltd.). Labelled DNA samples were mixed with 430 µl of 1x TE, pH 8.0 (Promega), transferred to YM-30 filters in a 1.5 ml microfuge tube and centrifuged at 8000 *x*g for 10 min. Then, the flow-through was discarded, 480 µl of 1x TE was added, and the tubes were centrifuged again at 8000 *x*g for 10 min. Next, the YM-30 filters were inverted into a fresh 1.5 ml microfuge tube and centrifuged at 8000 *x*g to collect the purified labelled sample. The volume of the purified eluate was measured and recorded. If the volume of the purified labelled DNA was >21 µl, it was returned to its filter, centrifuged at 8000 *x*g for 1 min and the flow through was discarded. Then the filters were re-inverted and the same steps were repeated until sample volumes were equal to or less than 21 µl. If the volume was <21 µl, 1x TE buffer (pH 8.0) was used to bring the volume up to 21 µl.

2.2.15.5 Labelling yield and specific activity determination

The labelling yield and the specific activity were determined using the Nanodrop. From the main menu of the ND-1000 software, the microarray measurement module was selected and the sample type was changed to DNA-50 (double-stranded DNA). 1x TE buffer (pH 8.0) was used to set the blank reading, and then 1.5 µl of purified labelled DNA was used for quantitation. The absorbance was recorded at A_{260nm} (DNA), A_{550nm} (cyanine 3) and A_{650nm} (cyanine 5). The specific activity was calculated with the following equation:

Specific activity (pmol dye/ µg DNA) = pmol per µl dye / µg per µl DNA

Next, equal volumes of the cyanine-3 labelled "normal" DNA was mixed with the cyanine-5 labelled "tumour" DNA.

2.2.15.6 Microarray hybridisation

Prior to hybridisation, the combined labelled DNA samples from the previous section were mixed with 50 µl cot-1 DNA (1.0 mg/ml) (Invitrogen Ltd., Renfrew) (to prevent the hybridisation of repetitive DNA sequences), 52 µl Agilent 10x blocking agent and 260 µl of Agilent Hi-RPM buffer, then incubated at 95°C for 3 min and at 37°C for 30 min. Then, the samples were centrifuged at 6000 *x*g for 1 min prior to hybridisation on the microarray slides. A clean gasket slide with 4 chambers was placed in an Agilent SureHyb chamber base with the gasket label facing up. Then, 100 µl of samples prepared as above were loaded into each chamber in a “drag and dispense” manner. Following that, the microarray (4x44K or 4x180K) was placed with the Agilent labelled barcode side (active side) facing down. The microarray slides were handled carefully without touching the edges. The SureHyb chamber cover was then placed onto the array-gasket sandwiched slides and they were clamped tightly. The assembled chamber was then rotated vertically to spread the samples, wet the microarray and to check the mobility of bubbles. Next, the assembled chamber was placed in the rotator rack of the hybridisation oven set at a temperature of 65°C and rotation speed of 20 RPM. The slides were left to hybridise for 24 hr.

2.2.15.7 Microarray washing

Wash procedure A of the Agilent’s protocol version 6 was used. Three separate 250 ml slide staining dishes were used. The first 2 slide staining dishes were filled with wash buffer 1 at room temperature and the 3rd was filled with wash buffer 2 pre-warmed overnight at 37°C. Slide staining dish number 2 was placed on a magnetic stir plate and a slide rack and a magnetic stir bar were placed inside it. Slide staining dish number 3 was left in a water bath set at 37°C. After the washing dishes were in place, the hybridisation chamber was removed from the oven and the bubbles inside were checked to confirm that they were still rotating freely in the array-gasket sandwich. Next, the clamp of the hybridisation chamber and the chamber cover were removed. The array-gasket sandwich was then taken out and quickly submerged in wash buffer 1 (in the first slide staining dish). The sandwich was opened from the barcode end using clean plastic forceps. The array slide was then placed in the slide rack in slide staining dish number 2 for 5 min. Afterwards, slide staining dish number 3 was taken out of the water bath and placed on a heated magnetic stir plate with a magnetic stir bar inside. Next, the slide rack was transferred to dish number 3 for 1 min. The slide rack was then removed slowly and the array slide was scanned immediately.

2.2.15.8 Microarray scanning

Microarray scanning was performed using an Agilent C scanner and scanner control software v8.3 (Agilent). The washed slides were placed in slide holders with the Agilent barcode facing up. The assembled slide holders were then placed into the scanner carousel. For the 4x180K arrays, Agilent G3_CGH was selected with 3 μm resolution while for the 4x44K arrays, Agilent HD_CGH profile was selected with 5 μm resolution. After checking the scanner status, the slides slots were selected and scanning was performed.

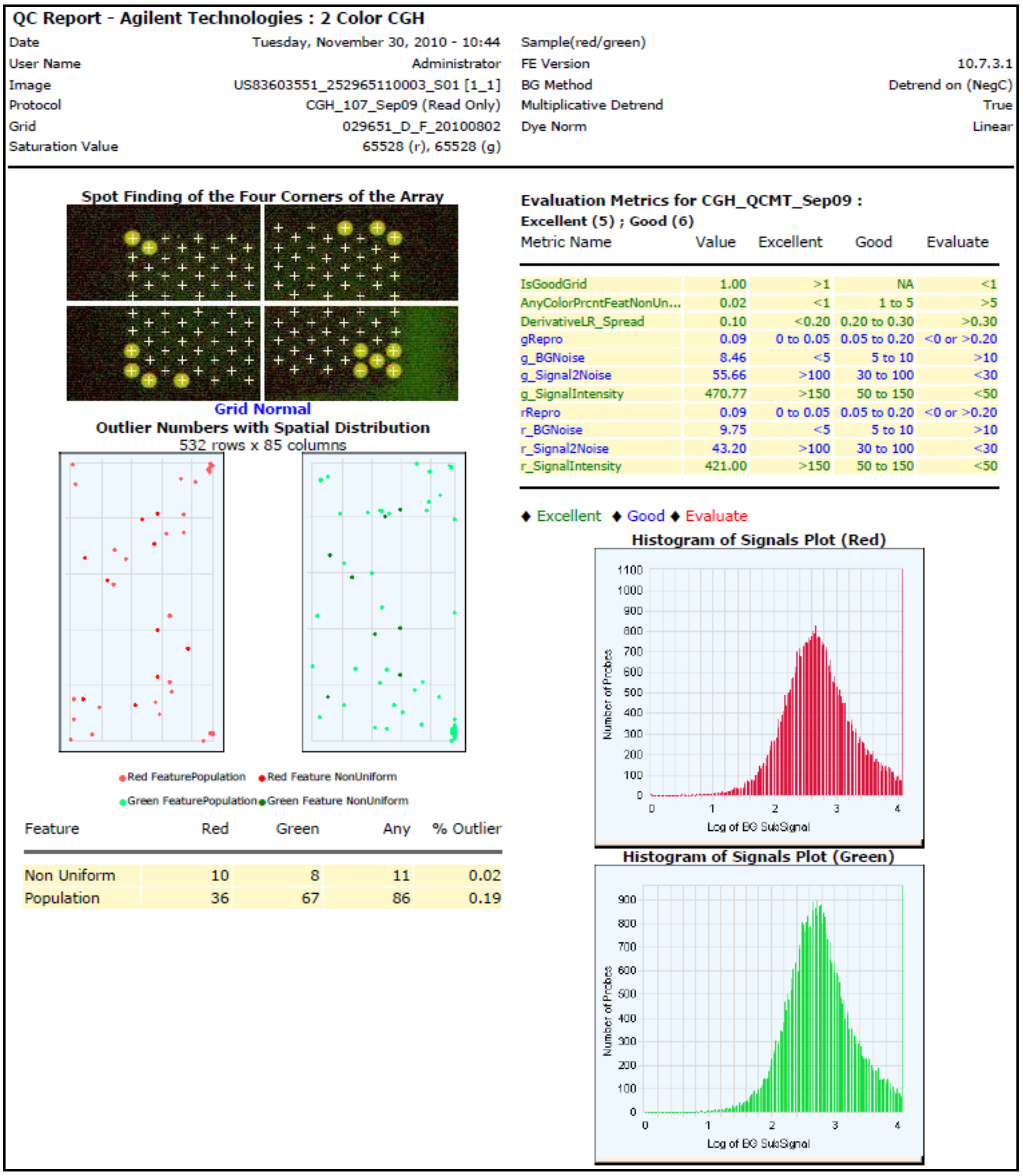
2.2.15.9 Data extraction using feature extraction software

The feature extraction software version 10.5.1.1 (Agilent) was used to extract the microarray TIFF images. The software was launched and an FE project was opened. Then using the “add new extraction set” icon, the .tif files of the scanned arrays were added. The grid template and the CGH FE protocol were automatically assigned depending on the scanned microarray platform. The FE project was then saved and the extraction was performed.

2.2.15.10 Quality control

In order to ensure the integrity and quality of the aCGH results, 11 QC metrics were closely monitored for each array (Figure 2.9). These QC metrics check the placement of the array GRID, the noise of the experiment determined by the Derivative Log

Ratio Spread (DLRS), the signal intensity and the signal background (green and red), the signal reproducibility of non-control replicate probes and the presence of any outliers. Each metric was scored as excellent, good or evaluate. "Evaluate" metrics are usually of low quality. Each array will have a final QC status of either pass, marginal or fail. Array QC metrics were closely monitored with special attention to DLRS values as they greatly affect reliable analysis.



QC Report - Agilent Technologies : 2 Color CGH

Date	Tuesday, November 30, 2010 - 10:44	Sample(red/green)	
User Name	Administrator	FE Version	10.7.3.1
Image	US83603551_252965110003_S01 [1_1]	BG Method	Detrend on (NegC)
Protocol	CGH_107_Sep09 (Read Only)	Multiplicative Detrend	True
Grid	029651_D_F_20100802	Dye Norm	Linear
Saturation Value	65528 (r), 65528 (g)		

Spot Finding of the Four Corners of the Array

Grid Normal

Outlier Numbers with Spatial Distribution

532 rows x 85 columns

Feature	Red	Green	Any	% Outlier
Non Uniform	10	8	11	0.02
Population	36	67	86	0.19

Evaluation Metrics for CGH_QCMT_Sep09 :
Excellent (5) ; Good (6)

Metric Name	Value	Excellent	Good	Evaluate
IsGoodGrid	1.00	>1	NA	<1
AnyColorPrcntFeatNonUn...	0.02	<1	1 to 5	>5
DerivativeLR_Spread	0.10	<0.20	0.20 to 0.30	>0.30
gRepro	0.09	0 to 0.05	0.05 to 0.20	<0 or >0.20
g_BGNoise	8.46	<5	5 to 10	>10
g_Signal2Noise	55.66	>100	30 to 100	<30
g_SignalIntensity	470.77	>150	50 to 150	<50
rRepro	0.09	0 to 0.05	0.05 to 0.20	<0 or >0.20
r_BGNoise	9.75	<5	5 to 10	>10
r_Signal2Noise	43.20	>100	30 to 100	<30
r_SignalIntensity	421.00	>150	50 to 150	<50

◆ Excellent ◆ Good ◆ Evaluate

Histogram of Signals Plot (Red)

Histogram of Signals Plot (Green)

Figure 2.9 An example of aCGH QC report

2.2.15.11 Data analysis

Genomic workbench software version 5.0.14 (Agilent) was used for the main analysis of the aCGH microarray results. The CGH module was selected and a new experiment was created. Then, the results files, exported from the feature extraction software, were imported and the analysis performed.

Several algorithms were available in the genomic workbench software v5.0 to perform the analysis. For defining and displaying the aberrations, 5 algorithms were available; circular binary segmentation (CBS), hidden Markov model (HMM), z-score based method, aberration detection method 1 (ADM1) and quality weighted interval score algorithm, also called the aberration detection method 2 (ADM2). In support for the aberration detection algorithms, data centring and correction algorithms were also available. These algorithms were used to increase the stringency of the aberration detection algorithms. Moreover, replicate probes and feature and aberration filters were also available to ensure the validity of the calls. Finally, 2 common aberration analysis algorithms were provided; the t-test based method and the modified context corrected analysis algorithm. The following sections aim to briefly describe the concepts, limitations and advantages of the available algorithms. Most of the information presented is based on the CGH interactive analysis user guide for the Agilent genomic workbench software (v5.0).

2.2.15.11.1 Aberration detection algorithms

CBS and HMM are used for genome segmentation and detection of aberrant probes respectively. CBS identifies genomic points where the mean $\log_2$ratio scores change between intervals. On the other hand, aberrant probes identified by HMM can be collectively used to define aberrant regions. CBS and HMM are not user friendly and their outputs require a lot of processing and input from the operator to identify the aberrant regions. The latter might cause subjective calls, thus, they were not used in this project. The z-score based method detects aberrations by searching for genomic regions enriched with probes that have $\log_2$ratios significantly different from a calculated baseline. The algorithm does so using a user defined sliding window of a fixed size. The disadvantage of the z-score algorithm is the fixed size of the sliding window which can limit the ability of the algorithm to identify some aberrations.

ADM1 searches for any genomic regions and intervals that consistently have $\log_2$ratio values that are statistically significantly different from the zero $\log_2$ratio. The level of significance is determined by a user-defined statistical threshold which is recommended by Agilent to be set to a value of 6.0 for reliable results. ADM1 also determines the optimal size and the breakpoints of the identified aberrations. ADM2 uses the same approach as ADM1, however, it also incorporates the quality information for the $\log_2$ratio values. Therefore, ADM2 deals better with lower quality or noisy data and is more stringent in identifying aberrations. ADM2 was the aberration detection method of choice in this project and its statistical threshold was set to 6.0 as recommended by Agilent. Another advantage of choosing ADM2 (or ADM1) is that the common aberration analysis algorithms (t-test based method and the modified context corrected analysis method) cannot deal with the outputs from the other aberration detection algorithms (CBS, HMM and z-score).

2.2.15.11.2 Data centring and correcting algorithms

Data centring and correcting algorithms were included in the analysis to increase the confidence in the called aberrations.

2.2.15.11.2.1 Centralisation

In general, the aberration detection algorithms will assume that the $\log_2$ratio values are centred around zero. That is, that the mean fluorescence values for the reference and the sample DNA are the same. This is often true for samples with stable or relatively stable genomes. However, for highly aberrant and unstable genomes (such as tumour genomes), the measured centre might deviate from the zero $\log_2$ratio and the above assumption can lead to many aberrations being falsely called. Centralisation is an additional normalisation algorithm used to re-centre the $\log_2$ratio values and ensure that the zero-point value actually reflects the most common ploidy status of the samples. The algorithm will find a constant value for each sample based on its genomic status to re-normalise the $\log_2$ratios. The default centralisation parameters (bin size= 10 and threshold 6.0) were used as recommended by Agilent. The use of this centralisation algorithm was shown to increase the accuracy of the aberration detection algorithms (Chen *et al.* 2008).

2.2.15.11.2.2 Fuzzy zero correction algorithm

ADM1 and 2 can sometimes call large aberrations with a very low $\log_2$ratio average. These aberrations might be of genuinely low amplitude because of tumour heterogeneity. However, they could also be false aberrations caused by noise, and only reaching statistical significance because of the large number of probes in the area. This is caused by the limited ability of the ADM (1&2) algorithms to estimate errors over long genomic intervals. When used, the Fuzzy zero correction algorithm will apply a global error model to all of the aberrations called by the ADM algorithms. The global error model will assume two sources of noise; a local probe to probe noise where the error is not correlated between different probes within the same interval and a global noise where the error is correlated between the different probes. This meticulous global error estimation approach will avoid erroneous calling of large intervals with a high level of noise.

2.2.15.11.3 Feature filter, aberration filter and replicate probes

Feature filters were used to exclude non-uniform and saturated signals from the green and red fluorescence channels based on information from the feature extraction software. The default feature filter settings were used as recommended by Agilent. Additionally, an aberration filter was used which requires a minimum of 3 probes for any aberration to be called. This will remove any aberrations called based on deviation of 2 probes or less from the normalised zero $\log_2$ratio. Thus, the remaining calls will be of higher confidence. Finally, replicate probes were included in all of the Agilent arrays. The values of the intra-array replicate probes were combined during analysis to increase the statistical power for aberration detection.

2.2.15.11.4 Common aberration analysis

For common aberration analysis, context corrected common aberration (COCA) algorithm was chosen over the t-test based method. The t-test based method looks for commonly aberrant regions regardless of the amplification/deletion nature of the aberration. On the other hand, the COCA algorithm considers amplifications and deletions separately. It uses the aberrant regions identified by ADM2 (or ADM1) to construct a set of candidate intervals. Then, it will generate a statistical score reflecting the significance of each aberration in each sample. This significance score will be corrected based on the aberration status of each specific sample. This correction can be made in a genomic context for all the aberrations or chromosomal context. Because of the nature of CRC DNA in which some of the chromosomes are highly aberrant and others are relatively stable, a chromosomal context approach was applied for correction.

The COCA algorithm will then test the hypothesis that each candidate aberration is common between any number of samples and it will generate a combined statistical score for each candidate. Candidate intervals will be reported as common aberrant regions if their combined statistical score (COCA score) was more than the recommended t-test p-value threshold of 0.05 and the overlap threshold of 0.9. The latter threshold means that for any common regions with > 90% overlap, the COCA algorithm will only report the one with higher statistical significance. COCA was previously shown to accurately identify common aberrant regions in cancer samples (Ben-Dor *et al.* 2007).

2.2.16 Cell lines culture

Human cell lines were cultured in DMEM medium (Lonza, Belgium) supplemented with 10% foetal calf serum (FCS) and 1% pencillin-streptomycin in a 37^{o}C incubator with 5% CO_2. Standard sterile cell culture techniques were used to maintain the cell lines, which were all performed in a microbiological grade 2 safety cabinet.

2.2.17 Western blotting

Western blotting is a technique that can is used to detect the presence and the amount of specific proteins in a cell lysate. Briefly, proteins are separated according to size using Sodium Dodecyl Sulfate Polyacrylamide Gel Electrophoresis (SDS-PAGE), transferred to a nitrocellulose membrane and then detected using a specific antibody. Western blotting was used in this project to specifically assess the NFKBIA and the phosphorylated RELA protein levels in cellular extracts of various cell lines (Section 2.1.4).

2.2.17.1 Protein extraction

Confluent tissue culture plates were always used for protein extraction. The DMEM medium was removed and the cells were washed twice with PBS. Trypsin:versene (1:1) mixture (Lonza) was then added to the cells and the plates were incubated for few minutes (time varied according to cell lines) at 37ºC. When the cells were completely detached, 10 ml of DMEM medium (without antibiotics) were added to the plate. After pipetting up and down, the cells/medium mixture was transferred to 10 ml sterile tubes and centrifuged at 1300 RPM for 3 min. The cell pellet was then washed twice with 1x PBS. After the second wash, the cell pellet was resuspended in 1 ml 1x PBS and transferred to 1.5 ml eppendorf tube and centrifuged for at 2100 RPM for 3 min at 4ºC. Meanwhile, 1 ml of 1x cell lysis buffer was prepared by mixing 100 µl of 10x cell lysis buffer (Cell Signalling), 100 µl protease inhibitor (Roche), 100 µl phosphatase inhibitor (Roche), 10 µl serine protease inhibitor, termed phenylmethanesulfonylfluoride (100mM PMSF) and 690 µl distilled water. The PBS was then fully removed and ~300 µl of the freshly prepared ice-cold 1x cell lysis buffer were added to the tubes (the volume varied according to pellet size). After thoroughly pipetting the pellet up and down, the extracts were passed through a fine-gauge needle ~30 times in order to release nuclear proteins. The tubes were incubated on ice for 30 min and then centrifuged at 14500 RPM for 5 min at 4ºC. The supernatant was then transferred to a new 1.5 ml tube and stored at -80ºC.

2.2.17.2 Protein quantification

Aliquots of the extracted protein samples were slowly defrosted on ice and then diluted in distilled water. Serial dilutions of 1 mg/ml BSA solution were also prepared in distilled water to be used as standards for protein quantification. The serial dilutions of both the samples and the BSA were then quantified using Bradford reagent (Bio-Rad Laboratories Ltd.) according to the manufacturer's instructions.

2.2.17.3 Sodium Dodecyl Sulfate Polyacrylamide Gel Electrophoresis

SDS denatures proteins and coats them in negative charges which will ensure their separation largely according to size during electrophoresis. SDS polyacrylamide gels were used to separate the protein extracts in preparation for Western blotting. Resolving and stacking gels were freshly prepared prior to each experiment using the reagents listed in Table 2.7 and Table 2.8 respectively. Tetramethylethylenediamine (TEMED) and 10% ammonium persulphate (APS) were only added when the gels were ready to be poured. The Bio-Rad mini protein gel apparatus was assembled according to the manufacturer's instructions (Bio-Rad Laboratories Ltd.) and 7 ml of the resolving gel was poured between the gel electrophoresis plates. Next, 200 µl of distilled water was carefully poured on top to eliminate air bubbles. Gels were left to set for ~15 min. After removing the water completely, the stacking gel was poured to fill the space between the plates and the well comb was inserted. The gel was ready to use after ~15 min. While the gel was polymerising, 20 µg of the protein samples were prepared by mixing with an equal volume of 2x Leammli buffer (5% v/v ß-mercaptoethanol) (Sigma-Aldrich) and boiling for ~3 min.

The electrophoresis plates were placed in the mini protein gel apparatus and 1x SDS buffer was poured into the central and the outer compartments. The well combs were then removed and 5 µl of Precision Plus Protein^{TM} Standard (Bio-Rad Laboratories Ltd.) was added into the first well. The protein samples were then added into the adjacent wells and the gels were electrophoresed for ~2 hr at 100 V.

Table 2.7 Reagents and quantities for SDS-PAGE resolving gels

	Quantity (ml)	
Resolving gel	10%	12%
Distilled water	5.30	4.00
30% polyacrylamide	6.70	8.00
10% SDS	0.20	0.20
1 M Tris-Cl (pH 8.8)	7.50	7.50
10% APS*	0.25	0.25
TEMED	0.025	0.025

*10% APS were freshly prepared by dissolving 0.1g ammonium persulphate in 1ml distilled water.

Table 2.8 Reagents and quantities for SDS-PAGE stacking gel

Stacking gel 5%	**Quantity (ml)**
Distilled water	6.80
30% polyacrylamide	1.70
1 M Tris-Cl (pH 6.8)	1.25
10% SDS	0.10
10% APS*	0.125
TEMED	0.005

*10% APS were freshly prepared by dissolving 0.1 g ammonium persulphate in 1ml distilled water.

2.2.17.4 Protein transfer

Once electrophoresis was complete, the gel was removed from the mini protein gel apparatus and the stacking gel was detached. The resolving gel was then incubated in transfer buffer for 30 min with shaking. The protein transfer apparatus (Bio-Rad Laboratories Ltd.) was assembled according to the manufacturer’s instructions. Whatman Nitrocellulose membrane (GE Healthcare UK Ltd., Little Chalfont) was cut according to the size of the gel and incubated in the transfer buffer for ~15 min. Sponges and 3 mm thick blotting filter papers (GE Healthcare UK Ltd.) were also prepared and briefly soaked in transfer buffer. The protein transfer sandwich was prepared by placing a wet sponge onto the cathode of the transfer apparatus, followed by 3 mm blotting filter paper, the gel, the nitrocellulose membrane, another 3 mm blotting paper, the second sponge and the anode (Figure 2.10). The protein sandwich was placed into the transfer apparatus and it was filled with transfer buffer. Transfer was carried out at 200 mA for 90 min.

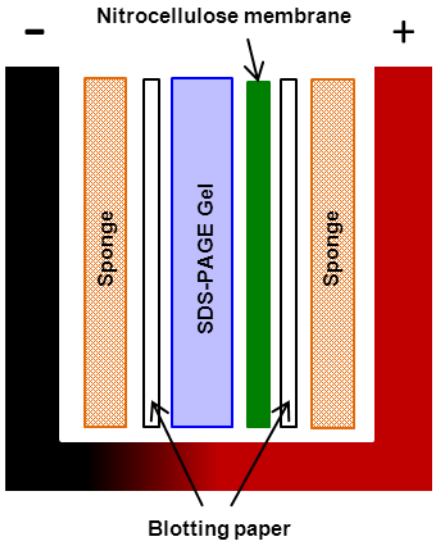


Figure 2.10 Western-blot protein transfer sandwich

The transfer apparatus was disassembled and the membrane was stained with Ponceau S solution (Sigma-Aldrich) to check the efficiency of the transfer. The membrane was washed with 1x TBS-T and placed in blocking buffer (5% dried milk in 1% TBS-T) for 2 hr at room temperature to prevent non-specific binding of the antibody. During the blocking step, fresh dilutions of the primary mouse anti-human antibody (NFKBIA or p-RELA) (Cell signalling) were prepared according to the manufacturer's instructions (1:1000 in the blocking buffer). The blocking buffer was removed and the membrane was incubated with 10 ml of the freshly diluted primary antibody overnight at 4ºC.

Next day, the membrane was washed twice for 10 min with 1x TBS-T on a shaker. Following that, mouse secondary antibody (Sigma-Aldrich) was added after dilution in 10 ml blocking buffer according to the manufacturer's recommended concentration (1:1000). The membrane was incubated for 1 hr at room temperature on shaker. Two 10 min 1X TBS-T washes were then performed and the membrane was transferred onto a dry and clean plastic plate. Enhanced Chemiluminescent (ECL) reagent (GE Healthcare UK Ltd.) was prepared by mixing an equal volume of reagent 1 and 2 and then added slowly to cover the membrane. The covered membrane was incubated for 2 min at room temperature and the ECL reagent was then poured off. The membrane was wrapped in a plastic film and analysed using the luminescent image analyser LAS-3000 (Fuji-film, Bedfordshire) according to manufacturer's recommendations. The membranes were then washed 3 times for 10 min with 1x TBS-T on a shaker and the incubated in blocking buffer (5% dried milk in 1% TBS-T) for 1 hr at room temperature and then stained with the reference protein anti-α-tubulin antibody (Sigma-Aldrich) using the same procedure described above.

2.2.18 siRNA transfection and NFKBIA knockdown

Small interfering RNA (siRNA) silencing is an effective mechanism that has been demonstrated to downregulate gene expression, and thus protein levels, in human cells (Lares *et al.* 2010). siRNA technology was used to temporarily knockdown

NFKBIA proteins by transfecting cells with a pool of *NFKBIA*-siRNA (Fisher Scientific UK Ltd.,) designed to complement *NFKBIA* messenger RNA (mRNA) (Table 2.9).

Table 2.9 *NFKBIA*-siRNA target sequences

***NFKBIA*-siRNA on target plus smart pool sequences**
GUGCUGAUGUCAAUGCUCA
AGGACGAGCUGCCCUAUGA
GCUGAUGUCAACAGAGUUA
AGUCAGAGUUCACGGAGUU

Transfection of the *NFKBIA*-siRNA into the target cells was performed using the transfection reagent Lipofectamine 2000 according to the manufacturer's instructions (Invitrogen Ltd.). On the day before siRNA transfection, 2-2.5x10^5 target cells were added to a 6 well cell culture plate containing 2 ml DMEM medium (with 10% FCS but without antibiotics) and incubated overnight at 37^{o}C with 5% CO$_2$. The next day, *NFKBIA*-siRNA, "negative" siRNA (designed not to complement any known human m-RNA sequence) (Eurogentec S.A., Southampton) and Lipofectamine 2000 were prepared. For each sample, 10 µl of 20 µM *NFKBIA*-siRNA (or negative siRNA) stock was mixed with 240 µl DMEM medium (without 10% FCS or antibiotics). Also, for each sample, 5 µl Lipofectamine 2000 was mixed with 245 µl DMEM medium (without 10% FCS or antibiotics). After 5 min incubation at room temperature, the 250 µl Lipofectamine 2000 in DMEM was gently mixed with the 250 µl *NFKBIA*-siRNA in DMEM (or "negative" siRNA). These mixtures were then incubated for 20 min at room temperature. Meanwhile, the plated cells were washed with 1x PBS and then 1.5 ml DMEM medium (with 10% FCS but without antibiotics) was added to each well. The Lipofectamine 2000/siRNA (*NFKBIA* or "negative") mix was added to the cells which were then incubated at 37^{o}C with 5% CO$_2$. siRNA knockdown efficiency was determined by Western blotting (Section 2.2.17).

2.2.19 MTT cell proliferation assay

In order to investigate the effect of NFKBIA knockdown on CRC cell lines growth, the number of viable transfected cells was estimated using the 3-(4,5-Dimethylthiazol-2-yl)-2,5-diphenyltetrazolium bromide (MTT) colorimetric assay. The MTT assay reflects the number of viable cells based on the activity of dehydrogenase enzymes (Mosmann 1983). The MTT assay was performed according to the manufacturer's instructions (ATCC). Briefly, 100 µl of 1x10^4 transfected cells were plated in 96 well plates and incubated (24-72 hr) at 37^{o}C with 5% CO_2. Next, 10 µl of the MTT reagent was added to the wells and incubated for 4 hr at 37^{o}C with 5% CO_2. Then, 100 µl of the solubilisation reagent (10% SDS, 10mM HCl) was added and incubated overnight at 37^{o}C with 5% CO_2. Absorbance was then measured at 570 nm using the MTT module on the Multiskan FC plate spectrophotometer (Thermo Scientific). The MTT test was performed on 3 consecutive days following transfection of cells with *NFKBIA*-siRNA and the "negative" siRNA.

2.2.20 Colony formation experiment

Colony forming activity is considered a representation of cell transformation which reflects the malignancy of cancer cells in culture (Bredel *et al.* 2010). Colony formation test was performed to examine the effect of the NFKBIA knockdown on the malignant behaviour of the transfected CRC cell lines. Briefly, 100 µl of 1x10^4 transfected cells were mixed with 8 ml DMEM medium (with 10% FCS but without antibiotics), plated in cell culture petri dishes and incubated at 37^{o}C with 5% CO_2. Colonies were observed up to 2 weeks and counted following staining with 0.4% methylene blue. The test was performed on cells transfected with *NFKBIA*-siRNA and the "negative" siRNA.

2.2.21 Statistical tests

2.2.21.1 Chi-square association test

The chi-square test (χ^2) was used to compare genotype frequencies for the 6bp ins/del polymorphism rs3834129 (and *CASP8* CNV23598) between CRC cases and controls. The different genotypes (Ins/Ins, Ins/Del and Del/Del) and phenotypes (such as cases and controls) were tabulated in a 2x3 contingency table (Table 2.10). The equation $x^2 = \sum \frac{|O-E|^2}{E}$ was used to calculate the χ^2 statistic, where *O* is observed genotype frequency and *E* is expected genotype frequency. Expected frequencies are calculated using the contingency table. For example, the expected frequency of cases to have an Ins/Ins genotype is calculated by: (T_{case} x $T_{ins/ins}$)/T (Table 2.10). To determine whether the χ^2 is significant or not, the degrees of freedom (df) of the test are calculated using the equation: df = (number of columns – 1) x (number of rows – 1). The df for 2x3 contingency tables thus equals 2. The chi-square distribution table can then be used to obtain the p-value for the χ^2 statistic. P-values <0.05 are considered to represent a significant result. The "CHIDIST" function in Microsoft Excel spreadsheet template was used to automatically determine the p-values for the various contingency tables.

Table 2.10 A 2x3 contingency table

	Genotype			
Phenotype	**Ins/Ins**	**Ins/Del**	**Del/Del**	**Total**
Case	x	y	z	T_{case}
Control	a	b	c	$T_{control}$
Total	$T_{ins/ins}$	$T_{ins/del}$	$T_{del/del}$	T

2.2.21.2 Hardy Weinberg Equilibrium

In conditions of HWE, allele and genotype frequencies remain constant in the population across generations. This is an indication of random mating in large populations in the absence of selective pressure and mutations. Under conditions of HWE, the genotype frequencies are given by:

$$p^2+2pq+q^2=1$$

Where p and q stand for the frequencies of the common and the rare alleles respectively. Therefore, the expected genotype frequencies of the population in HWE were calculated as follows:

E for the homozygous common= p^2 x n

E for the homozygous rare= q^2 x n

E for the heterozygoutes = 2pq x n

Where "n" is the total number of genotyped samples in each assay.

Testing for HWE is an essential quality control step in association studies. Departure from HWE might indicate low assay quality and genotyping errors. It might also indicate problems with the genotyped cohort such as small sample size, the presence of selection or population stratification. The chi-squared test with df=1 was used to check for any deviations from HWE. If the P-value was <0.05, the population genotypes were considered to be inconsistent with HWE.

2.2.21.3 Odds ratios

The odds ratio (OR) is an estimate of relative risk of CRC in a subset of patients with a specific genotype (e.g. Ins/Del) compared to another subset with a different genotype (e.g. Ins/Ins). The OR was calculated by dividing the odds in one group (e.g. Ins/Del) by the odds in the other group (Ins/Ins). An additive model was also used to calculate the OR based on the assumption that the risk of CRC increases additively with each copy of the rare allele (Del). OR and confidence intervals applied in this thesis were estimated by Dr. Angela Cox using Stata (v9).

2.2.21.4 Fisher's exact test

Fisher's exact test (2x2 contingency table) was used to examine the significance of association between various genetic and/or clinical variables. The test was performed using an online Fisher's exact test calculator available on Microsoft research website (http://research.microsoft.com, accessed May – September, 2011)

2.2.21.5 Mann Whitney U test

Mann Whitney U test is applied to assess if two independent sets of observations have statistically similar or different values. The test was used in this project to compare the number of CNA between CIMP-L/N and CIMP-H samples. The Mann Whitney U test was calculated using an online Microsoft Excel spreadsheet template (www.holah.karoo.net/Mann-Whitney%20U-test.xls, accessed May 2011).

2.2.21.6 Spearman's correlation test

Spearman's correlation test was used to examine the strength of relationship between two different genetic defects (variables). The Spearman correlation coefficient (r_s) was calculated using the CORREL function available in Microsoft Excel spreadsheet. A negative correlation coefficient value indicates an inverse relationship between the two tested variables.

2.2.22 Splice site prediction

In order to predict the physiological relevance of the novel intronic sequence alterations identified in this study, online splice-site prediction tools were used to predict any potential effects of the sequence alterations on splicing. Five online tools were applied on each of the sequence alterations. Several online tools were used to overcome shortcomings and weaknesses in any of the tools, as recommended by Houdayer *et al.* (2008). The splice site prediction tools web addresses are listed below (the default parameters were maintained unless specified):

1) Splice site prediction by neural network:

   http://fruitfly.org:9005/seq_tools/splice.html.

2) Exonic Splicing Enhancer (ESE) Finder :

   http://rulai.cshl.edu/cgi-bin/tools/ESE3/esefinder.cgi.

   Parameters used:

a) Matrix Library: Splice sites

b) Matrix: 5’ splice sites donor of human

3) MaxEntScan: scorsplice (MES):

http://genes.mit.edu/burgelab/maxent/Xmaxentscan_scoreseq.html.

All the scoring models available on the tool were chosen.

4) Relative Enhancer and Silencer Classification by Unanimous Enrichment (RESCUE-ESE): http://genes.mit.edu/burgelab/rescue-ese/.

5) NetGene2 splice prediction: http://www.cbs.dtu.dk/services/NetGene2/

(All the above web tools were accessed July, 2011).

2.2.23 Sequence nomenclature

Throughout this study, the Human Genome Variation Society (HGVS) recommendations were followed for sequence nomenclature (http://www.hgvs.org/, accessed May, 2009). The coding DNA sequences were numbered from the 1st nucleotide of the start codon as the +1 position and they were prefixed with a “c”. The amino acid (aa) sequences were numbered from the 1st amino acid as the +1 position, standard three letter amino acid codes were used and each aa sequence was prefixed with a “p”.

3. Caspase 8 Inherited Variants and CRC Predisposition

3.1 Introduction

As described in Section 1.3, CRC is a major cause of cancer related deaths worldwide (Parkin *et al.* 2005, Jemal *et al.* 2011). The prognosis of CRC cases is largely dependent on the presenting cancer stage at the time of diagnosis. More than 90% of the patients diagnosed in Duke's stage A survive for ~5 years compared to only ~5% when diagnosed in Duke's stage D (de la Chapelle 2004). Therefore, early diagnosis of CRC can play an important role in increasing the survival of patients.

CRC is divided into sporadic and hereditary cases with the latter accounting for ~25% of the incidence (Lichtenstein *et al.* 2000, de la Chapelle 2004, Broderick *et al.* 2007). Environmental factors, such as diet and physical activity, have a major impact on CRC risk (Jong *et al.* 2002), however, genetics also play a key role in CRC predisposition with an estimated contribution of ~35%, placing it 2nd among common cancers in terms of heritability (Lichtenstein *et al.* 2000). Nonetheless, highly penetrant and well characterised mutations are rare and account for ~5% of all CRC cases and ~20% of the hereditary CRC cases (Jong *et al.* 2002, de la Chapelle 2004). The majority of the hereditary cases and a large proportion of the sporadic CRC cases are probably caused by more common, but less penetrant, genetic variants and polymorphisms (Broderick *et al.* 2007). These are generally difficult to identify, since linkage studies do not have enough power to detect low penetrance variants. Nevertheless, through the use of SNP-based genetic association studies (based on both, candidate genes and genome wide) with large numbers of CRC cases and controls, some of these less penetrant genetic variants have already been identified (Jong *et al.* 2002, de la Chapelle 2004, Broderick *et al.* 2007, Haiman *et al.* 2007, Sun *et al.* 2007, Tomlinson *et al.* 2007, Zanke *et al.* 2007, Jaeger *et al.* 2008, Tenesa *et al.* 2008, Tomlinson *et al.* 2008). In general, all the identified common risk variants can only account for ~6% of the unknown CRC heritability (Lascorz *et al.* 2010). Identifying more of these lower penetrance genetic variants will play an important role in increasing our knowledge of the genetic pathways involved in CRC development with possible applications in the clinic (de la Chapelle 2004, Lerman and Shields 2004). Recently, it has been suggested that some of the missing heritability may be accounted for by copy number variants (CNVs) (Pearson and Manolio 2008). CNVs are structural copy number polymorphisms; insertions, deletions and duplications of more than 1Kb in size that occur across the human genome. CNVs have been considered as a form of genetic variation and they affect the risk of genetic diseases (Beckmann *et al.* 2007). Therefore, they can provide an alternative strategy to identify cancer predisposing genes (Venkatachalam *et al.* 2010).

Apoptosis, is a fundamental biochemical pathway that plays an essential role in normal tissue homeostasis and differentiation, as it helps to eliminate abnormal cells (Reed 2000, Fadeel and Orrenius 2005). One of the main hallmarks of cancer cells is their ability to evade apoptosis (Hanahan and Weinberg 2000) which is usually acquired by disrupting apoptotic genes (Butler *et al.* 1999). Caspase 8 is one of the initiator caspases that plays a major role in activating apoptosis. Several studies have shown the loss of caspase 8 expression in tumours such as neuroblastoma, medulloblastoma and small cell lung cancer (Shivapurkar *et al.* 2002b, Pingoud-Meier *et al.* 2003b, Stupack *et al.* 2006). Moreover, inactivating *CASP8* mutations were previously reported in several cancer types such as CRC and gastric cancer (Shivapurkar *et al.* 2002b, Kim *et al.* 2003, Soung *et al.* 2005). Additionally, *CASP8* SNPs such as D203H have been shown to affect breast cancer risk (Cox *et al.* 2007).

Recently, a large association study in a Chinese population identified a 6bp in/del (rs3834129) in the *CASP8* promoter as a risk factor for multiple cancers including

CRC (Sun *et al.* 2007). However, these results were not confirmed in European and multi-ethnic American populations (Haiman *et al.* 2008, Pittman *et al.* 2008). In an attempt to clarify this apparent disparity, the initial aim of the work described in this chapter was to investigate the relationship between rs3834129 and CRC risk in a Caucasian population. Secondly, our lab had some preliminary data indicating an association between *CASP8* SNPs and CRC risk (Curtin *et al.* manuscript in preparation). Therefore, we decided to try to identify relatively rare and novel functional SNPs in *CASP8* that might affect CRC risk. This was performed by re-sequencing the coding region, intron/exon boundaries, the promoter region and the 3'UTR of *CASP8*. Rare variants could easily be missed by GWAS, because they are mainly based on a tagging SNP approach which uses specific known SNPs as representatives of multiple other common SNPs that occur together in strong linkage disequilibrium. Thus, rare variants are usually poorly captured by this approach. Finally, we also decided to develop an assay to investigate a possible role of *CASP8* CNV23598 in CRC predisposition. At the time this project was designed, only two CNVs were reported to occur in *CASP8*; CNV23598 and CNV23081. Both CNVs were not validated and their frequencies were unknown. However, we decided to develop an assay for investigating CNV23598 mainly because of its physical proximity to D302H, a *CASP8* variant identified by our group to affect breast cancer risk (Cox *et al.* 2007).

3.2 Results

3.2.1 rs3834129 genotyping

The SNP rs3834129 was genotyped by TaqMan assay (Section 2.2.9.1) (Figure 3.1) in 1193 cases and 1388 controls from the UK (Sheffield, Leeds and Dundee) and the USA (Utah) populations (these are described in Section 2.1.3). A summary some of the characteristics of the cases and controls included in the study is presented in Table 3.1.

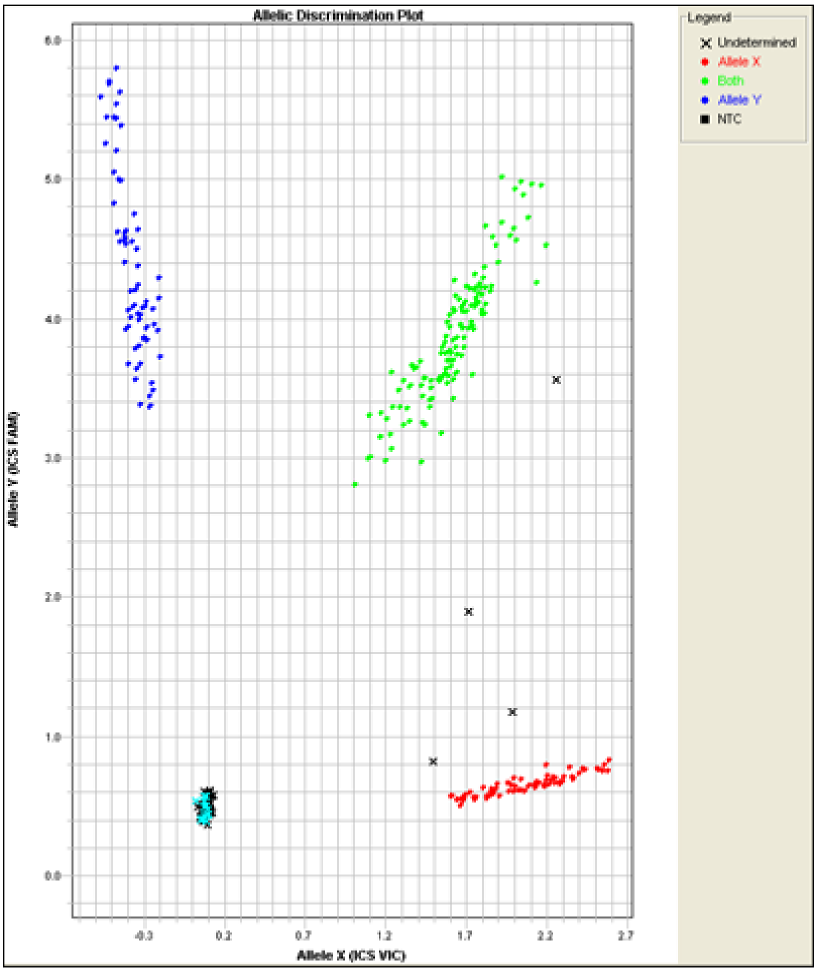


Figure 3.1 TaqMan genotyping

Genotyping clusters of rs3834129 Taqman assay from a 384 well plate from the Sheffield population. The X and the Y axes represent fluorescent signals from VIC® and FAM^{TM} labelled probes respectively. Blue and Red clusters represent homozygous insertion and homozygous deletion genotypes respectively. Green cluster represents heterozygous genotypes. The crosses represent undetermined samples.

Table 3.1 Summary of the cases and controls for rs3834129 genotyping

		Sheffield		Leeds		Dundee		Utah	
		Cases	Controls	Cases	Controls	Cases	Controls	Cases	Controls
		n (%)							
	Total	475 (100.0)	447 (100.0)	270 (100.0)	227 (100.0)	137(100.0)	365(100.0)	455 (100.0)	449 (100.0)
Sex	Male	264 (55.6)	221 (49.4)	153 (56.7)	131 (57.7)	85(62.0)	189 (51.8)	250 (55.0)	250 (55.7)
	Female	211 (44.4)	220 (49.2)	116 (43.0)	96 (42.3)	52(38.0)	173 (47.4)	205 (45.0)	199 (44.3)
	Unknown	NA	6 (13.4)	1 (0.4)	NA	NA	3 (0.82)	NA	NA
Family History*	None	393 (82.7)	405 (90.6)	228 (84.4)	202 (89.0)	109 (79.6)	315 (86.3)	62 (13.6)	420 (93.5)
	1 relative	69 (14.5)	36 (8.1)	37 (13.7)	24 (10.6)	26 (19.0)	45 (12.3)	264 (58.0)	25 (5.6)
	≥ 2 relatives	13 (2.7)	6 (1.3)	5 (1.9)	1 (0.4)	6 (4.4)	5 (1.4)	129 (28.4)	4 (0.9)
Age**	≤50	25 (5.3)	18 (4.0)	11 (4.1)	11 (4.9)	6 (4.4)	16 (4.4)	54 (11.8)	NA
	51-59	82 (17.3)	98 (21.9)	40 (14.8)	29 (12.8)	20 (14.6)	67 (18.5)	69 (15.2)	NA
	60-69	127 (27.6)	171 (38.3)	94 (34.8)	85 (37.4)	51 (37.2)	143 (39.5)	134 (29.5)	NA
	≥70	237(49.9)	160 (35.8)	124 (45.9)	102 (44.9)	60 (43.8)	136 (37.6)	185 (40.7)	NA
	Mean (Range)	68 (30-96)	66 (26-86)	67.4 (45-89)	67.1 (45-80)	67.2 (46-80)	66.7 (45-86)	64.9 (29-91)	NA
CRC site***	Proximal	122 (25.7)	NA	57 (24.4)	NA	22 (16.0)	NA	158 (32.6)	NA
	Distal	126 (26.5)	NA	81 (34.6)	NA	49 (35.8)	NA	147 (30.4)	NA
	Rectal	190 (40.0)	NA	92 (39.3)	NA	62 (45.3)	NA	115 (25.3)	NA
	Unknown	37 (7.8)	NA	4 (1.7)	NA	4 (2.9)	NA	64 (14.1)	NA

* Family history includes 1st degree relatives only.

** Controls for Utah were samples collected from previous studies. All controls were cancer free and matched by sex and year of birth (Age will reflect the age at recruitment for previous studies, therefore it will not necessarily match for this study (Section 2.1.3.3).

*** CRC site for Utah includes 29 cases with multiple primary CRC.

3.2.1.1 Quality control

In order to assess the quality of the DNA samples and the accuracy of the TaqMan assay and genotyping results, several quality control assessments were performed. The assay genotyping call rate represents the number of successfully genotyped samples out of the total number of samples. It gives an indication of both assay and sample quality. Usually, a call rate of ~95% indicates that both the assay and the samples are of high quality. The call rate achieved for the different genotyped populations was between 92-96%, as shown in Table 3.2. In each 384 well taqman plate, 5-10% of the samples were duplicated. The duplicate concordance represents the percentage of the duplicated samples that were successfully called with the same genotype. A duplicate concordance of 98% represents high accuracy. An overall duplicate rate of 97.5% was obtained for all of the genotyped cohorts combined (range 95-100%) (Table 3.2). As mentioned in Section 2.2.21.2, testing for Hardy

Weinberg equilibrium (HWE) is an important step to investigate the quality of the genotyping assay. Table 3.2 summarises the p-values for the populations genotyped for rs3834129. The chi-squared test was performed on the control genotypes only. All the p-values were >0.05 suggesting that all the genotyped populations were consistent with HWE. In summary, the quality control analysis of the genotyping results indicates reasonable data quality.

Table 3.2 Quality control results of the genotyped cohorts

Study Group		Samples (n)	Call rate (%)	Overall call rate (%)	Duplicate Concordance % (n)*	HWE (p-value)
Sheffield	**Cases**	436	92.0	92.3	94.7 (94)	0.507
	Controls	442	92.5			
Leeds	**Cases**	262	94.7	96.1	100.0 (41)	0.254
	Controls	226	97.8			
Dundee	**Cases**	136	94.8	94.0	97.2 (36)	0.990
	Controls	364	93.7			
Utah	**Cases**	451	92.0	93.3	100.0 (71)	0.714
	Controls	444	94.6			

3.2.1.2 Allelic discrimination and genotypes

Table 3.3 represents a summary of the rs3834129 genotype counts for the cases and controls from the genotyped populations. Moreover, Table 3.4 shows the observed allele frequencies in this study, together with those from previous publications and dbSNP. Comparison with the known frequencies of rs3834129 shows an agreement between our results and the results from the European and multi-ethnic American populations. However, it shows substantial difference from the Asian population.

Table 3.3 rs3834129 Genotyping results

		Genotyping results			
		n (%)			
		Ins/Ins	**Ins/Del**	**Del/Del**	**Total**
Sheffield	**Cases**	107 (26.7)	186 (46.4)	108 (26.9)	401
	Controls	119 (29.1)	197 (48.2)	93 (22.7)	409
Leeds	**Cases**	53 (21.4)	126 (50.8)	69 (27.8)	248
	Controls	58 (26.2)	102 (46.2)	61 (27.6)	221
Dundee	**Cases**	40 (31.0)	60 (46.5)	29 (22.5)	129
	Controls	80 (23.5)	171 (50.1)	90 (26.4)	341
Utah	**Cases**	107 (25.8)	210 (50.6)	98 (23.6)	415
	Controls	116 (27.6)	213 (50.7)	91 (21.7)	420

Table 3.3 summarises the genotyping results for rs3834129.

Table 3.4 rs3834129 frequencies

	rs3834129 allele frequencies	
	Insertion	**Deletion**
Sheffield	0.53	0.47
Leeds	0.49	0.51
Dundee	0.49	0.51
Utah	0.53	0.47
Pittman *et al.* 2008	0.50	0.50
Haiman *et al.* 2008	0.50-0.52	0.48-0.50
European panel dbSNP(133)	0.452	0.548
Sun *et al.* 2007	0.76	0.24
Asian panel dbSNP(133)	0.783	0.217

Pittman *et al.* 2008 and Haiman *et al.* 2008 studies were performed in European and multi-ethnic American populations respectively. Sun *et al.* 2007 study was performed in a Chinese population. dbSNP allele frequency was obtained from the database release 133.

3.2.1.3 Association of rs3834129 with colon cancer

A contingency table chi-squared test (Degree of freedom (df) =2) was used to investigate whether there was a significant difference in the frequency of the genotypes between cases and controls (Section 2.2.21.1). The test was applied separately on the four different populations using observed and expected values based on the null hypothesis of no association between genotype and case status. The p-values for the 4 populations are summarised in Table 3.5.

The odds ratios (OR) for CRC relative risk of the heterozygous and the homozygous deletion (the rare allele) genotypes were calculated in comparison to the homozygous insertion (the common reference allele) (Section 2.2.21.3). OR were calculated separately for the different populations and a combined OR was also computed for the 4 populations with adjustment for study size. Finally, an additive OR model assuming an increase in CRC risk with each copy of the deletion allele (allele dose) was calculated for each study separately and with all the studies combined. The calculated OR are summarised in Table 3.5. All the calculated p-values were > 0.05 and all the estimated OR 95% confidence intervals overlapped 1.0. Therefore, these data provide no evidence of association between rs3834129 and CRC risk.

Table 3.5 Chi-squared test p-values and odds ratios

		Odds ratios (OR)		
Study	**p-value***	**Heterozygous OR (95% CI)**	**Homozygous Del OR (95% CI)**	**Additive model OR (95% CI)**
Sheffield	0.369	1.050 (0.756-1.459)	1.292 (0.882-1.890)	1.134 (0.937-1.372)
Leeds	0.429	1.352 (0.858-2.130)	1.238 (0.745-2.056)	1.105 (0.857-1.424)
Dundee	0.234	0.702 (0.434-1.134)	0.644 (0.366-1.134)	0.797 (0.599-1.061)
Utah	0.736	1.069 (0.773-1.479)	1.168 (0.792-1.721)	1.080 (0.890-1.310)
Combined data**		1.040 (0.860-1.258)	1.110 (0.891-1.384)	1.053 (0.944-1.176)

Table 3.5 summarises the chi-squared test p-values, and the odd ratios for the different populations used in rs3834129 genotyping.

* 2x3 Chi-squared p-values testing the null hypothesis that there is no significant difference in genotypes frequencies between cases and controls.

** Combined OR were adjusted for the sizes of the 4 studies based on a logistic regression analysis by Dr Cox.

3.2.2 Caspase 8 gene sequencing

In order to try and identify any rare and possible coding variants that might have an effect on caspase 8 function and CRC risk, the coding region of the *CASP8* gene, the intron/exon boundaries, the promoter region and the 3'UTR (Figure 3.2) were sequenced in the peripheral blood DNA samples of 94 random CRC cases from the Sheffield population sample described in Section 2.1.3.1 (Table 3.6).

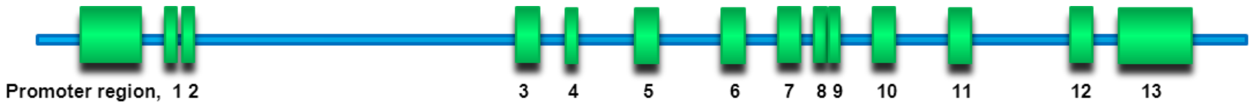


Figure 3.2 *CASP8* promoter region and exons

A schematic representation of *CASP8* promoter region and exons sequenced in the project.

Table 3.6 Patients and tumour characteristics

Median age (range), years	67 (30-90)
Gender	**n (%)**
Male	51 (54.3)
Female	43 (45.7)
Tumour location	**n (%)**
Proximal	31 (33.0)
Distal	27 (28.7)
Rectal	34 (36.2)
Unknown	2 (2.1)
Duke's Stage	**n (%)**
A	7 (7.4)
B	26 (27.7)
C	44 (46.8)
D	3 (3.2)
Unknown	14 (14.9)
Family history CRC	**n (%)**
None	72 (76.6)
1 relative	11 (11.7)
≥ 2 relatives	11 (11.7)
Metastastatic	**n (%)**
Yes	7 (7.4)
No	83 (88.3)
Unknown	4 (4.3)
Differentiation	**n (%)**
Well	7 (7.4)
Well/Moderate	53 (56.4)
Moderate	20 (21.3)
Poor	4 (4.3)
Unknown	7 (7.4)

Table 3.6 summarises the patients and tumour characteristics of the 94 sequenced CRC cases. Family history was defined by 1st degree relatives having CRC.

3.2.2.1 Novel sequence variants

Six novel variants were identified in the screened cohort. To date, none of these variants have been previously reported. All the novel variants (except c.646+25) were confirmed by bi-directional sequencing and are summarised in Table 3.7.

Table 3.7 Novel sequence variants

Variant	Location	Genomic position*	Number of cases
-565 G>A	Promoter region	202097619	2
c.1-8338Del115	Exon 3	202122873-202122988	1
c.1-7982A>G	Intron 3	202123228	1
c.646+25A>G	Intron 8	202137524	1
c.1488+388DelATTA	Intron 13	202151702-202151705	1
c.1488+1163C>T	Intron 13	202152480	1

* Genomic positions are according to human genome build 37.2.

3.2.2.1.1 Promoter region variant

The first variant was identified in the promoter region of *CASP8* in two different CRC cases (C201 and M202) (Figure 3.3). The variant is a G to A transition at nucleotide position -565 (-565 G>A) from the transcription initiation site in exon 1. The *CASP8* promoter region is well characterised and the regulatory sequences have been previously mapped and functionally validated (Figure 3.4) (Liedtke *et al.* 2003). The novel -565 variant does not occur within any of the known functional sites. Moreover, *in silico* analysis using the online tool PROMO for the identification of putative transcription factors binding sites (TFBS) did not predict the introduction of any new sites (http://alggen.lsi.upc.es/cgi-bin/promo_v3/promo/promoinit.cgi?dirDB=TF_8.3, accessed July, 2011). Finally, the genomic position 202097619 was checked using the UCSC genome browser (http://genome.ucsc.edu, assembly Feb 2009, accessed July, 2011) for evolutionary conservation (across; Rhesus, Mouse, Dog, Elephant and Opossum) and the presence of regulatory regions assessed by DNase hypersensitivity clusters and histone marks. The genomic position was not highly conserved and it did not occur within any regulatory region.

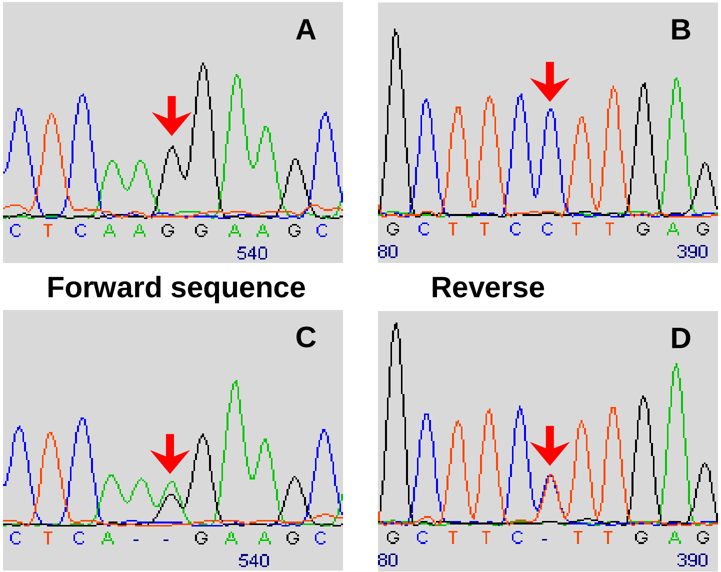
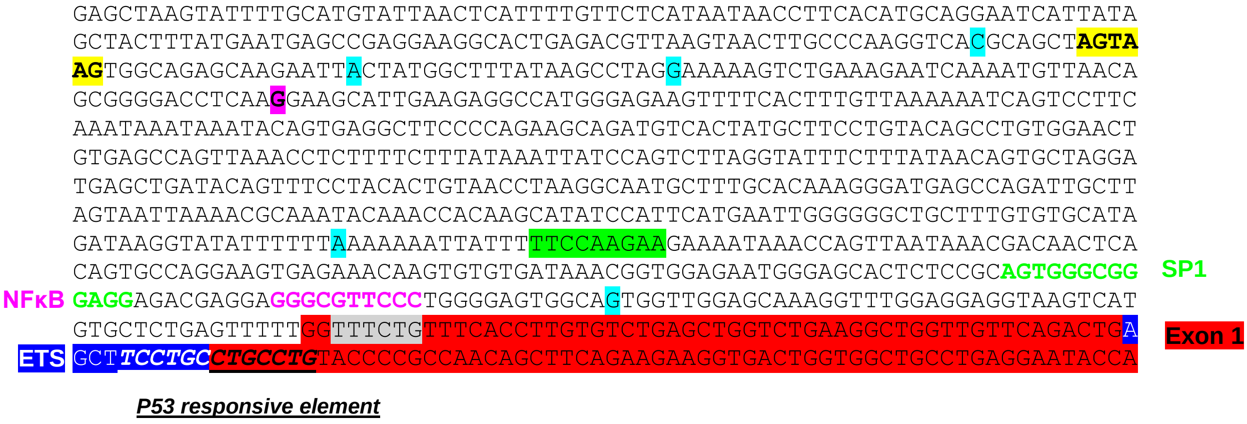


Figure 3.3 The novel variant -565 G>A

Bidirectional sequencing of the promoter region showing the -565 G>A variant in a wild type sample (A and B) and C201 (C and D). The site of the variant is marked with a red arrow.



```
     GAGCTAAGTATTTTGCATGTATTAACTCATTTTGTTCTCATAATAACCTTCACATGCAGGAATCATTATA
     GCTACTTTATGAATGAGCCGAGGAAGGCACTGAGACGTTAAGTAACTTGCCCAAGGTCACGCAGCTAGTA
     AGTGGCAGAGCAAGAATTACTATGGCTTTATAAGCCTAGGAAAAAGTCTGAAAGAATCAAAATGTTAACA
     GCGGGGACCTCAAGGAAGCATTGAAGAGGCCATGGGAGAAGTTTTCACTTTGTTAAAAAATCAGTCCTTC
     AAATAAATAAATACAGTGAGGCTTCCCCAGAAGCAGATGTCACTATGCTTCCTGTACAGCCTGTGGAACT
     GTGAGCCAGTTAAACCTCTTTTCTTTATAAATTATCCAGTCTTAGGTATTTCTTTATAACAGTGCTAGGA
     TGAGCTGATACAGTTTCCTACACTGTAACCTAAGGCAATGCTTTGCACAAAGGGATGAGCCAGATTGCTT
     AGTAATTAAAACGCAAATACAAACCACAAGCATATCCATTCATGAATTGGGGGGCTGCTTTGTGTGCATA
     GATAAGGTATATTTTTTTAAAAAAATTATTTTTCCAAGAAGAAAATAAACCAGTTAATAAACGACAACTCA
     CAGTGCCAGGAAGTGAGAAACAAGTGTGTGATAAACGGTGGAGAATGGGAGCACTCTCCGCAGTGGGCGG   SP1
NFκB GAGGAGACGAGGAGGGCGTTCCCTGGGGAGTGGCAGTGGTTGGAGCAAAGGTTTGGAGGAGGTAAGTCAT
     GTGCTCTGAGTTTTTGGTTTCTGTTTCACCTTGTGTCTGAGCTGGTCTGAAGGCTGGTTGTTCAGACTGA   Exon 1
ETS  GCTTCCTGCCTGCCTGTACCCCGCCAACAGCTTCAGAAGAAGGTGACTGGTGGCTGCCTGAGGAATACCA
```

P53 responsive element

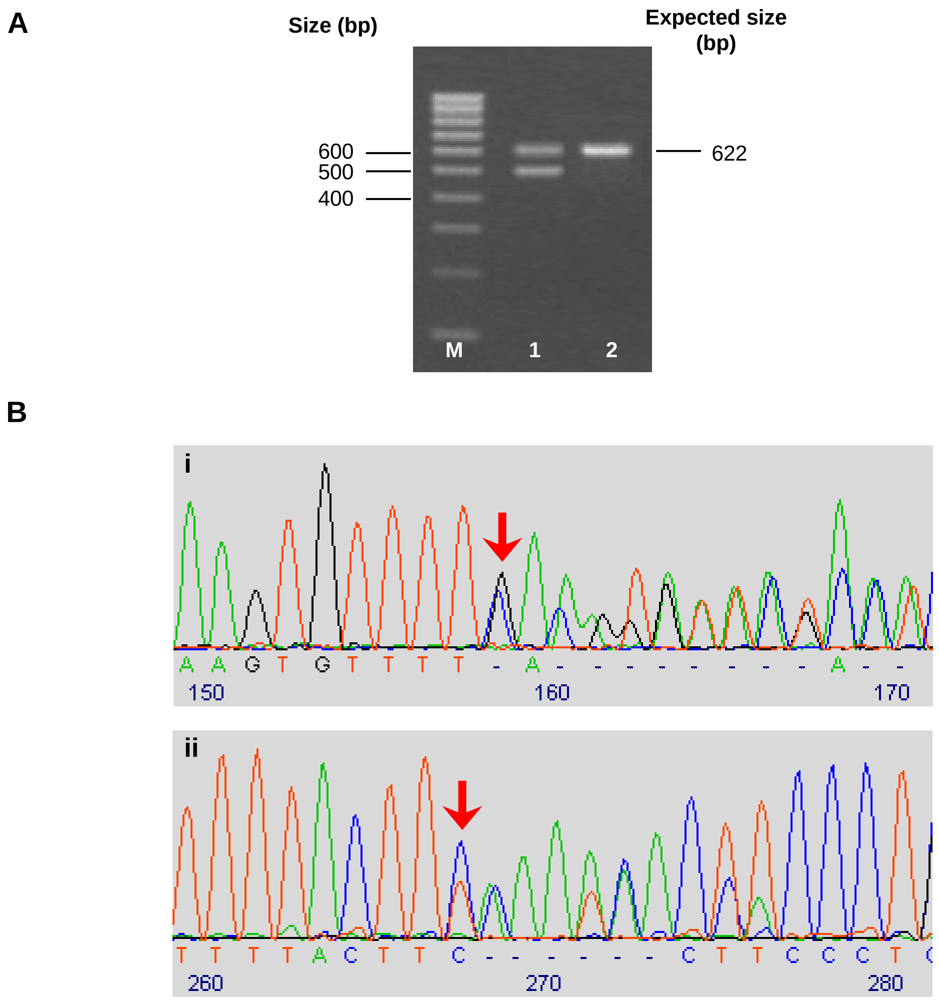
Figure 3.4 *CASP8* promoter region

The novel variant is highlighted in pink. Nucleotides highlighted in yellow and green represent the 6bp ins/del (rs3834129) and STAT1 binding site respectively. Nucleotides highlighted in turquoise represent known SNPs.

3.2.2.1.2 Exon 3 variant

The second variant was a 115bp deletion in exon 3 (c.1-8338Del115). The variant was identified in one CRC case only (C536) (Figure 3.5). Exon 3 is only expressed in the longest isoform of *CASP8* (Isoform G). c.1-8338Del115 results in an in-frame deletion of the first 11 amino acids (MEGGRRARVVI) of the isoform G precursor. The deleted 11 amino acids are not part of the known functional domains of caspase 8.

In silico analysis using the database of protein domains, families and functional sites Prosite (http://www.expasy.ch/prosite/ accessed July, 2011) did not predict any highly conserved or functional domains. Nevertheless, c.1-8338Del115 genomic region was checked using the UCSC genome browser (http://genome.ucsc.edu, assembly Feb 2009, accessed July, 2011) for evolutionary conservation (across; Rhesus, Mouse, Dog, Elephant and Opossum) and the presence of regulatory regions assessed by DNase hypersensitivity clusters and histone marks. The region was not highly conserved, however, the presence of strong DNase hypersensitivity cluster and histone marks were predicted in *CASP8* exon3 and the surrounding region (including the novel variant c.1-7982A>G described in the following section). However, *CASP8* exon 3 is only expressed in *CASP8* isoform G, which is not a predominantly expressed *CASP8* isoform (Scaffidi *et al.* 1997).



C

```
GCACGTCCATGAATTGTCTGCCACATCCCTCTTCTGAATGGTTGGAAATTGGGCATCTGTTCCTTTAAAC
AGGAAACATTTCTTGTTCGAGTGAGTCATCTCTGTTCTGCTTTAGGAGTAAAGTTTACCCTGCAGTTCCT
TCTGTGGTGAAGTTTTCTCTTTCTCTCGGAGACCAGATTCTGCCTTTCTGCTGGAGGGAAGTGTTTTCAC
AGGTTCTCCTCCTTTTATCTTTTGTGTTTTTTTTCAAGCCCTGCTGAATTTGCTAGTCAACTCAACAGGA
AGTGAGGCCATGGAGGGAGGCAGAAGAGCCAGGGTGGTTATTGAAAGTAAAAGAAACTTCTTCCTGGGAG
CCTTTCCCACCCCCTTCCCTGCTGAGCACGTGGAGTTAGGCAGGTTAGGGGACTCGGAGACTGCGATGGT
GCCAGGAAAGGGTGGAGCGGGTGAGTGCCTGTTGCCAAGGTGGCCTCTTCAACAGGAAACCACAATATTT
TTGTTTCTTGACTTGCTCTAGAAACAGGGCTGTGGGGGTGGGGAAGCAACTTGGATCTGCCCTTCTGAGG
ACACCTCTGGGTGCTGCCTGGCCCAGGTCTCCTGTGTGGTTTCTCTCTGAGCCGTTGCCTCTGACTTTGC
```

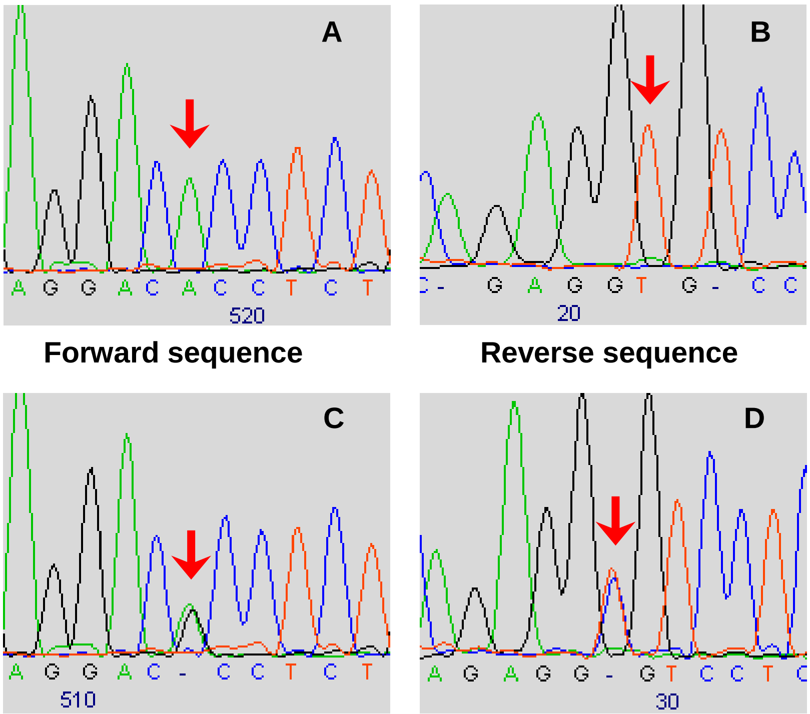
Figure 3.5 ***CASP8* exon 3 115 bp deletion**

A) Exon 3 deletion on 1.5% agarose gel (M: Hyperladder IV, 1: C536, 2: M225). B) Electropherogram of C536 showing c.1-8338Del115. (i and ii forward and reverse sequences respectively). Red arrows mark deletion breakpoints. C) *CASP8* exon 3 is marked in red. c.1-8338Del115 is highlighted in yellow and SNPs in turquoise. The underlined red nucleotides represent primers (*italics* for reverse).

3.2.2.1.3 Intronic variants

Four novel intronic variants were identified; c.1-7982A>G (intron 3) in case C371 (Figure 3.6), c.646+25A>G (intron 8) in case C109 (Figure 3.7) and c.1488+388DelATTA (intron 13) in control CO201 (used in the optimisation of the PCR conditions) (Figure 3.8) and c.1488+1163C>T (intron 13) in case C249 (Figure 3.9).

In *silico* analysis of the 4 intronic variants was performed using 5 different online splicing prediction tools as described in Section 2.2.21.4 and recommended by Houdayer *et al.* (2008). None of the tools used predicted any effect on splicing. Moreover, the genomic locations for the 4 intronic variants were checked using the UCSC genome browser (http://genome.ucsc.edu, assembly Feb 2009, accessed July 2011) for evolutionary conservation (across; Rhesus, Mouse, Dog, Elephant and Opossum) and the presence of regulatory regions assessed by DNase hypersensitivity clusters and histone marks. In terms of evolutionary conservation, c.1488+1163C>T was the only variant occurring in a highly conserved position (Rhesus, Mouse Dog and Elephant), however, it was not predicted to be in a regulatory region. The rest of the genomic variants (except for intron 3 c.1-7982A>G as explained in the previous section) did not occur within any regulatory regions.



Forward sequence

Reverse sequence

Figure 3.6 ***CASP8* intron 3 c.1-7982A>G**

Bidirectional sequencing of intron 3 region showing the variant c.1-7982A>G in a wild type sample (A and B) and C371 (C and D). The site of the variant is marked with a red arrow.

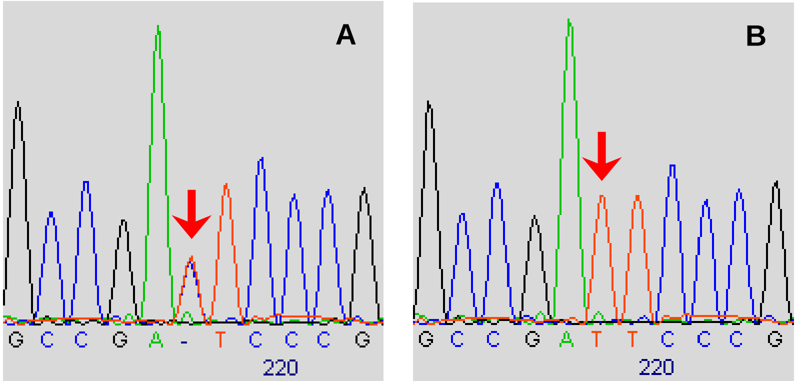
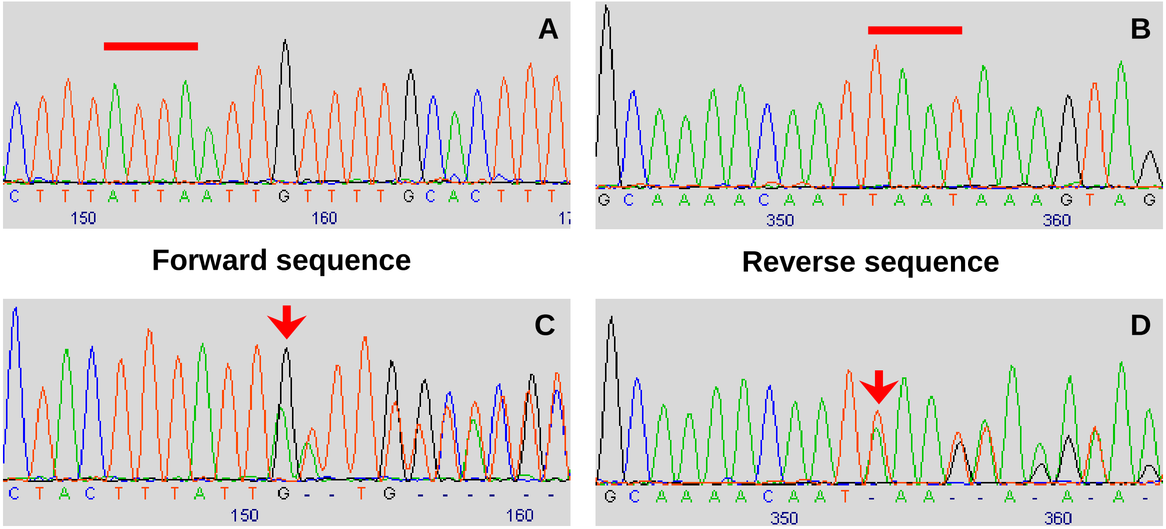


Figure 3.7 ***CASP8* intron 8 c.646+25A>G**

Reverse sequencing of intron 8 region showing the variant c.646+25A>G in a C109 (A) and a wild type sample (B). The site of the variant is marked with a red arrow.

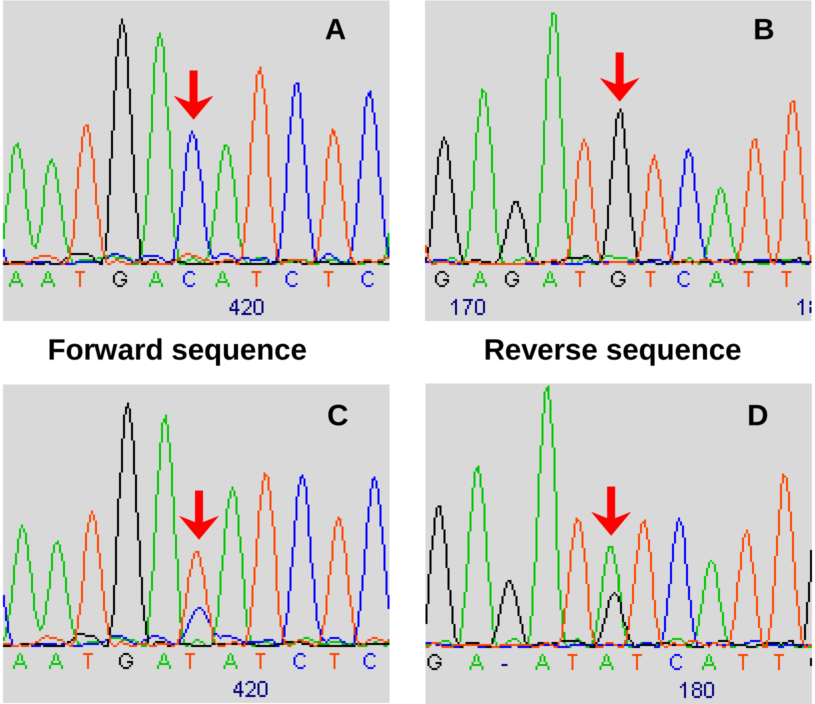


Forward sequence

Reverse sequence

Figure 3.8 ***CASP8*** **intron 13 c.1488+388DelATTA**

Bidirectional sequencing of intron 13 region showing the 4bp in/del (c.1488+388DelATTA) in a wild type sample (A and B) and CO201 (C and D). The breakpoints are marked with red arrows and the in/del sequence is marked with a horizontal red bar.



Forward sequence

Reverse sequence

Figure 3.9 ***CASP8*** **intron 13 c.1488+1163C>T**

Bidirectional sequencing of intron 13 region showing the variant c.1488+1163C>T in a wild type sample (A and B) and C249 (C and D). The site of the variant is marked with a red arrow.

3.2.2.2 Known polymorphisms

A summary of all the previously reported SNPs that were found in the course of this project is presented in Table 3.8. The frequencies of the SNPs in the sequenced cohort are compared to those from the European population panel (EGP-CEPH) available on dbSNP133 (http://www.ncbi.nlm.nih.gov/projects/SNP/). As shown in the table, the frequencies were similar and no major differences were observed.

Table 3.8 Known SNPs identified and their observed/known frequencies

SNP Reference Number	*CASP8* Exon/Intron	Alleles	Known Frequency* (db SNP 133)	Alleles Screened in the project (n)	Observed Frequency
rs3729647	Promoter	G/C	0.548/0.452	184	0.570/0.430
rs34224214	Promoter	C/T	0.976/0.024	182	0.940/0.060
rs3834129	Promoter	AGTAAG/-	0.452/0.548	188	0.569/0.431
rs35093568	Promoter	A/G	0.977/0.023	184	0.989/0.011
rs6747918	Promoter	G/A	0.568/0.432	184	0.490/0.510
rs36028425	Promoter	A/T	0.977/0.023	106	0.915/0.085
rs17860416	Promoter	G/A	0.932/0.068	188	0.952/0.048
rs3769823	Exon 3	G/A	0.705/0.295	188	0.680/0.320
rs3769824	Exon 3	T/C	0.979/0.021	188	0.963/0.037
rs2293554	Intron 5	T/G	0.952/0.048	188	0.930/0.070
rs13401994	Intron 5	G/T	0.977/0.023	188	0.973/0.027
rs36043647	Intron 11	T/C	0.932/0.068	188	0.936/0.064
rs1045485	Exon 12	G/C	0.773/0.227	188	0.872/0.128
rs1045487	Exon 12	G/A	0.977/0.023	188	0.968/0.032
rs34750049	Intron 12	C/T	0.977/0.023	188	0.984/0.016
rs41309820	Exon 13	C/T	Unknown	42	0.976/0.024
rs34841024	Exon 13	T/G	0.773/0.227	184	0.870/0.130
rs3769818	Exon 13	G/A	0.750/0.250	184	0.790/0.210
rs41309822	Exon 13	G/A	Unknown	184	0.990/0.001
rs17860428	Exon 13	G/A	0.841/0.159	182	0.841/0.159
rs3185378	Exon 13	C/G	0.682/0.318	180	0.611/0.389
rs2141331	Exon 13	C/T	0.932/0.068	180	0.906/0.094
rs35419671	Exon 13	C/T	0.977/0.023	186	0.994/0.006
rs17860432	Exon 13	G/C	0.977/0.023	184	0.935/0.065
rs17860433	Exon 13	A/G	0.955/0.045	184	0.973/0.027
rs1045494	Exon 13	T/C	0.977/0.023	184	0.967/0.033
rs13113	Exon 13	A/T	0.432/0.568	186	0.414/0.586
rs34461625	Exon 13	-/AT	0.977/0.023	186	0.973/0.027
rs1035140	Exon 13	A/T	0.500/0.500	184	0.544/0.456
rs35010052	Exon 13	-/A	0.955/0.045	184	0.956/0.044

* Known frequencies were obtained from the EGP-CEPH European panel available on NCBI website (dbSNP133).

3.2.3 Copy number variant CNV23598 genotyping

As of August 2009, when we started this investigation, 8410 CNV loci had been reported in the human genome (http://projects.tcag.ca/variation/). At the time, only two CNVs were present within *CASP8*; CNV23081 and CNV23598. Both of these CNVs were not independently confirmed and their frequencies were unknown. CNV23081 is located between intron 9 and intron 11 and CNV23598 is located in intron 11 downstream from CNV23081 (Figure 3.10). Because of the physical proximity to the D302H variant identified as a risk factor in breast cancer (Cox *et al.* 2007) (Figure 3.10), we wanted to investigate the association between CNV23598 and colorectal and breast cancer risk. To this end, we set up genotyping assays for CNV23598 in a set of CRC cases (n=189) and controls (n=95) from the Sheffield population sample (Section 2.1.3.1). We also made use of 47 carefully selected patients from a breast cancer study. The chosen patients included 24 cases homozygous for a *CASP8* breast cancer risk haplotype and 23 controls homozygous for a *CASP8* breast cancer protective haplotype (Shephard *et al.* 2009).

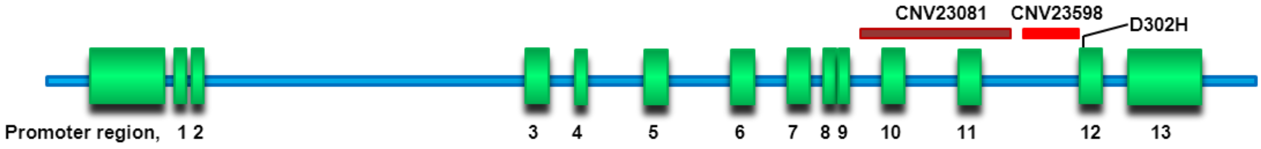


Figure 3.10 CNV23081 and CNV23598 in *CASP8*

A schematic representation of *CASP8* promoter region and coding sequence. The approximate location of CNV23081 and CNV23598 is presented by coloured lines. The breast cancer risk variant (D302H) location is also marked.

3.2.3.1 Optimisation of GAP PCR

Gel based GAP PCR, as described in Section 2.2.6.3 was initially used to investigate CNV23598 in CRC cases and controls from the Sheffield population. Three primers were designed to amplify both insertions and deletions in a duplex PCR reaction (Figure 3.11). The primers were optimised initially in two separate PCR reactions to give single amplification products. Subsequently, the primers were combined in a duplex reaction at equal concentrations of each of the 3 primers. However, preferential amplification of the smaller deletion region resulted in a stronger product (Figure 3.12 A). In order to obtain equal amplification of both products, a range of lower concentrations of the deletion reverse primer were used until products of similar intensity were obtained (Figure 3.12 B).

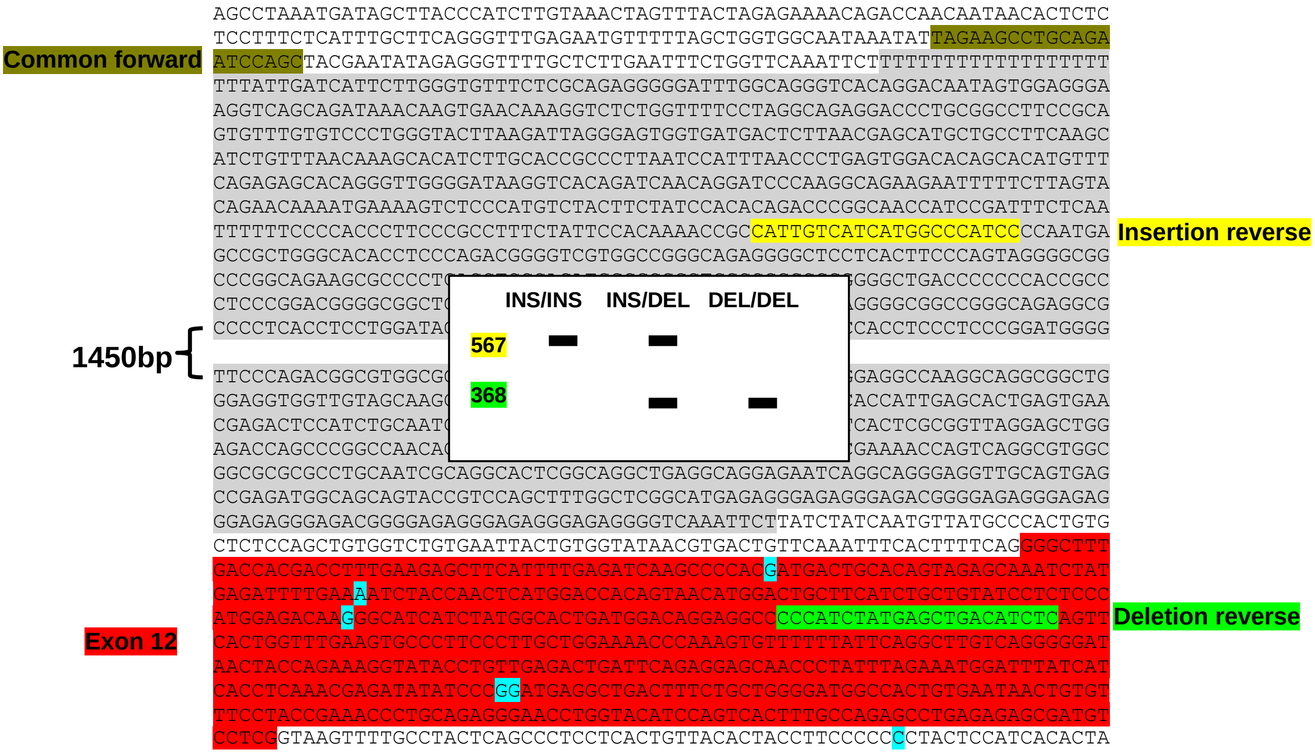

Common forward
Insertion reverse
1450bp
Deletion reverse
Exon 12
INS/INS
INS/DEL
DEL/DEL
567
368
AGCCTAAATGATAGCTTACCCATCTTGTAAACTAGTTTACTAGAGAAAACAGACCAACAATAACACTCTC
TCCTTTCTCATTTGCTTCAGGGTTTGAGAATGTTTTTAGCTGGTGGCAATAAATATTAGAAGCCTGCAGA
ATCCAGCTACGAATATAGAGGGTTTTGCTCTTGAATTTCTGGTTCAAATTCTTTTTTTTTTTTTTTTTT
TTTATTGATCATTCTTGGGTGTTTCTCGCAGAGGGGGATTTGGCAGGGTCACAGGACAATAGTGGAGGGA
AGGTCAGCAGATAAACAAGTGAACAAAGGTCTCTGGTTTTCCTAGGCAGAGGACCCTGCGGCCTTCCGCA
GTGTTTGTGTCCCTGGGTACTTAAGATTAGGGAGTGGTGATGACTCTTAACGAGCATGCTGCCTTCAAGC
ATCTGTTTAACAAAGCACATCTTGCACCGCCCTTAATCCATTTAACCCTGAGTGGACACAGCACATGTTT
CAGAGAGCACAGGGTTGGGGATAAGGTCACAGATCAACAGGATCCCAAGGCAGAAGAATTTTTCTTAGTA
CAGAACAAAATGAAAAGTCTCCCATGTCTACTTCTATCCACACAGACCCGGCAACCATCCGATTTCTCAA
TTTTTTCCCCACCCTTCCCGCCTTTCTATTCCACAAAACCGCCATTGTCATCATGGCCCATCCCCAATGA
GCCGCTGGGCACACCTCCCAGACGGGGTCGTGGCCGGGCAGAGGGGCTCCTCACTTCCCAGTAGGGGCGG
CCCGGCAGAAGCGCCCCTC GGGCTGACCCCCCCACCGCC
CTCCCGGACGGGGCGGCTC AGGGGCGGCCGGGCAGAGGCG
CCCCTCACCTCCTGGATAC CACCTCCCTCCCGGATGGGG
TTCCCAGACGGCGTGGCGC GGAGGCCAAGGCAGGCGGCTG
GGAGGTGGTTGTAGCAAGC ACCATTGAGCACTGAGTGAA
CGAGACTCCATCTGCAATC CACTCGCGGTTAGGAGCTGG
AGACCAGCCCGGCCAACAC GAAAACCAGTCAGGCGTGGC
GGCGCGCGCCTGCAATCGCAGGCACTCGGCAGGCTGAGGCAGGAGAATCAGGCAGGGAGGTTGCAGTGAG
CCGAGATGGCAGCAGTACCGTCCAGCTTTGGCTCGGCATGAGAGGGAGAGGGAGACGGGGAGAGGGAGAG
GGAGAGGGAGACGGGGAGAGGGAGAGGGAGAGGGGTCAAATTCTTATCTATCAATGTTATGCCCACTGTG
CTCTCCAGCTGTGGTCTGTGAATTACTGTGGTATAACGTGACTGTTCAAATTTCACTTTTCAGGGGCTTT
GACCACGACCTTTGAAGAGCTTCATTTTGAGATCAAGCCCCACGATGACTGCACAGTAGAGCAAATCTAT
GAGATTTTGAAAATCTACCAACTCATGGACCACAGTAACATGGACTGCTTCATCTGCTGTATCCTCTCCC
ATGGAGACAAGGGCATCATCTATGGCACTGATGGACAGGAGGCCCCCATCTATGAGCTGACATCTCAGTT
CACTGGTTTGAAGTGCCCTTCCCTTGCTGGAAAACCCAAAGTGTTTTTTATTCAGGCTTGTCAGGGGGAT
AACTACCAGAAAGGTATACCTGTTGAGACTGATTCAGAGGAGCAACCCTATTTAGAAATGGATTTATCAT
CACCTCAAACGAGATATATCCCGGATGAGGCTGACTTTCTGCTGGGGATGGCCACTGTGAATAACTGTGT
TTCCTACCGAAACCCTGCAGAGGGAACCTGGTACATCCAGTCACTTTGCCAGAGCCTGAGAGAGCGATGT
CCTCGGTAAGTTTTGCCTACTCAGCCCTCCTCACTGTTACACTACCTTCCCCCCCTACTCCATCACACTA

Figure 3.11 CNV23598 GAP PCR

CNV23598 is highlighted in grey, exon 12 in red and SNPs in turquoise. Primers are highlighted in different colours as indicated. The sizes of the expected products and the genotypes are represented in the box.

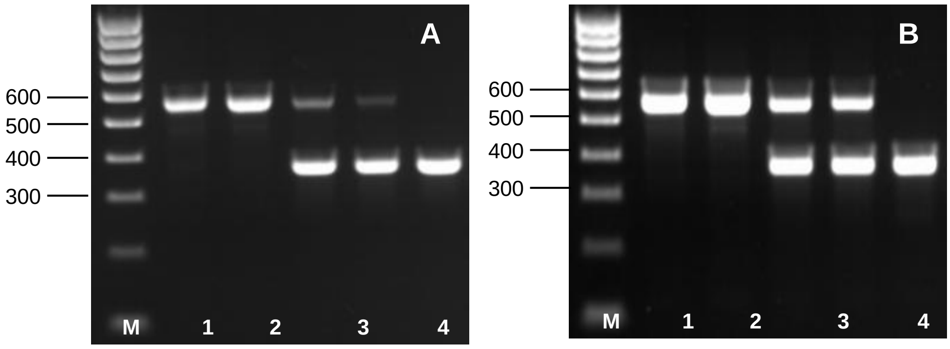


Figure 3.12 CNV23598 GAP PCR optimisation

Equimolar concentrations of the 3 primers resulting in preferential amplification of the smaller deletion product. B) Lower concentration of reverse insertion resulting in insertion and deletion products of similar intensities M) Hyperladder IV, 1&2) C003 & C011 (Ins/Ins), 3&4) M026 & C017 (Ins/Del), 5) C059 (Del/Del).

3.2.3.2 CNV23598 testing in breast cancer cases and controls

The optimised GAP PCR test was then used by Marina Parry (a PhD student in our research group) to test the 24 breast cancer cases and 23 controls. The results showed that 23 out of the 24 cases homozygous for the breast cancer *CASP8* risk haplotype were also homozygous for the insertion allele. On the other hand, the 23 controls homozygous for the protective haplotype were either heterozygotes (n=11) or homozygotes for the deletion allele (n=12). These results suggest a strong association between the insertion allele and the breast cancer risk haplotype (Table 3.9).

Table 3.9 CNV23598 genotypes for breast cancer cases and controls

	Homo INS	Heterozygous	Homo DEL
Cases	23	1	0
Controls	0	11	12

3.2.3.3 CNV23598 testing in CRC cases and controls

The assay was then used to genotype 96 non-metastatic CRC cases, 96 metastatic CRC cases and 96 controls from the Sheffield population sample. Table 3.10 summarises the genotyping results.

Table 3.10 Observed frequencies of CNV23598

	Genotype n (%)			
Total	**Ins/Ins**	**Ins/Del**	**Del/Del**	**Total**
Non-metastatic	21 (21.9)	64 (66.7)	11 (11.5)	96
Metastatic	10 (10.8)	58 (62.4)	25 (26.9)	93
Controls	12 (12.6)	60 (63.2)	23 (24.2)	95
Totals	43	182	59	284

As mentioned in Section 2.2.21.2, testing for HWE is a useful step in investigating the quality of the genotyping assay. Therefore, the HWE test was performed on the 95 successfully amplified controls in order to test the quality of the GAP PCR genotyping assay. The HWE chi-squared test p-value of 0.0063 was less than 0.05, suggesting that the genotyped data were inconsistent with HWE. We noticed that there was an excess of heterozygotes. The HWE test was then also applied to the non-metastatic and metastatic cases and the HWE chi-squared test p-values were 0.00065 and 0.0068 respectively. Therefore, the tested cohort might be non-random, or the GAP PCR test might not be accurate. The former seems unlikely since the samples are a subset of cases and controls demonstrated to be consistent with HWE (Section 3.2.1.1). It was observed that the genotype frequencies in the metastatic cases looked similar to the controls and both differed from the non-metastatic cases. However, as the populations are inconsistent with HWE, further investigations were carried out to confirm the validity and accuracy of the genotyping assay.

3.2.3.4 Comparison of fluorescent GAP PCR and gel-based GAP PCR and DNA sequencing of PCR products

In order to genotype a larger cohort of CRC cases for CNV23598, a high throughput technique was required. Fluorescent GAP PCR was optimised using a new set of reverse primers for the insertion and the deletion (Figure 3.13). The use of these new primers might also help in resolving the apparent genotyping issues described in the previous section. The insertion reverse was labelled with a FAM^{TM} dye and the deletion reverse was labelled with a VIC® dye. The PCR products for the 3 genotypes are shown in Figure 3.14. After optimising the fluorescent GAP PCR assay, 16 CRC cases (6 homozygous insertion, 8 heterozygotes and 2 homozygous deletion as determined by GAP PCR, Table 3.11) were genotyped using the fluorescent GAP PCR and the gel based GAP PCR for comparison. Furthermore, the GAP PCR products were sequenced to confirm their specificity towards the correct region of the genome.

Fluorescent GAP PCR and gel-based GAP PCR were concordant in the 16 samples. Moreover, the sequencing results confirmed the genotypes for all of the samples. Examples are shown (Figure 3.15 - Figure 3.17). Seven of the samples with the insertion sequence had a longer stretch of Ts at the beginning of the CNV23598 compared to the published sequence. This was confirmed by DNA sequencing and fluorescent GAP PCR (Figure 3.18).

Table 3.11 The 16 samples for CNV23598 confirmation

Sample	Genotype*
C004	Ins/Ins
C009	Ins/Ins
C010	Ins/Del
C017	Ins/Ins
C021	Ins/Ins
C047	Ins/Ins
C066	Ins/Del
C069	Ins/Del
C091	Ins/Ins
C099	Ins/Del
C106	Del/Del
C111	Ins/Del
C115	Ins/Del
C117	Ins/Del
C119	Del/Del
C121	Ins/Del

* CNV23598 genotypes in these samples were based on the gel-based GAP PCR results (Section 3.2.3.3).

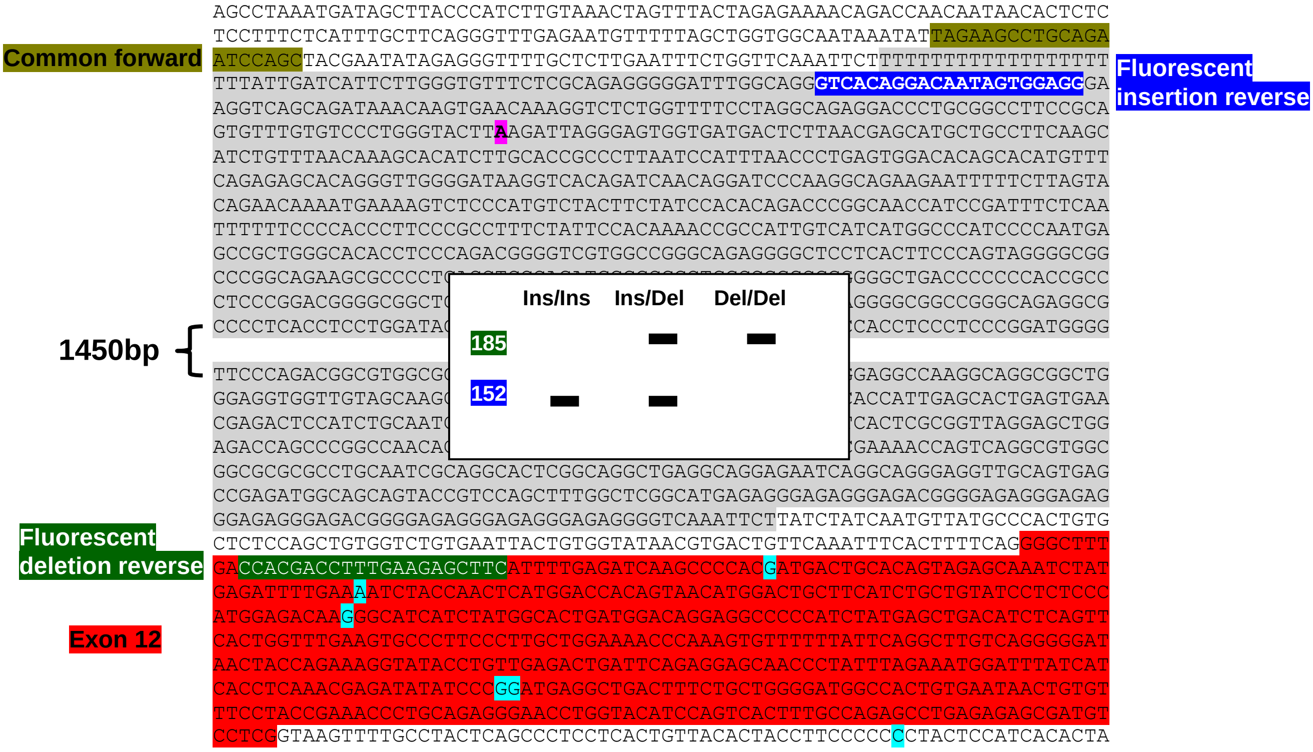


Figure 3.13 CNV23598 fluorescent PCR

CNV23598 is highlighted in grey, exon 12 in red and SNPs in turquoise. Fluorescent primers are highlighted in different colours as indicated. The c.898+5147A/G SNP is highlighted in pink. The sizes of the expected products and the genotypes are represented in the box.

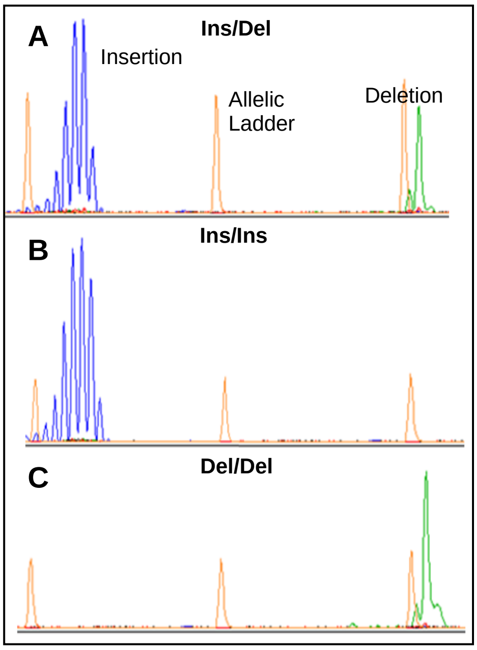


Figure 3.14 The 3 genotypes of CNV23598 using QF-PCR.

The multiple peaks of the insertion product are caused by Taq slippage due to the stretch of Ts at 5' end of the insertion sequence (Figure 3.13).

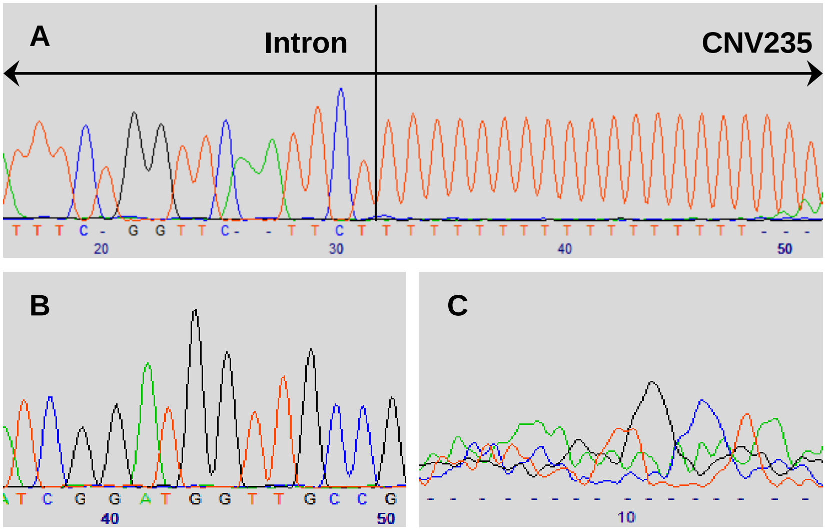


Figure 3.15 Homozygous insertion CNV23598

DNA sequencing of sample C017 with homozygous Insertion genotype using the primers designed for the GAP PCR. A) Common forward, the figure indicates the start of CNV23598, B) Insertion reverse confirming CNV23598 sequence, C) Deletion reverse failed to generate PCR product.

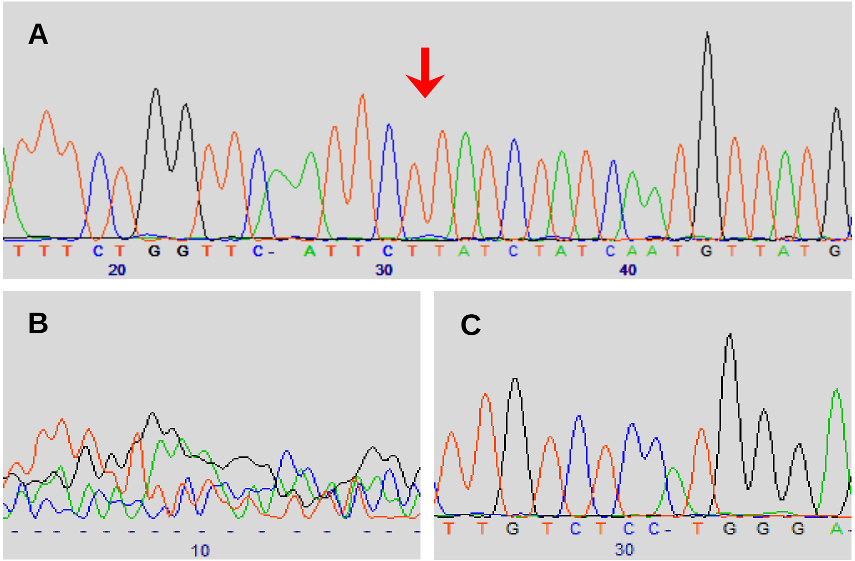


Figure 3.16 Homozygous deletion CNV23598

DNA sequencing of sample C119 with homozygous deletion genotype using the primers designed for the GAP PCR. A) Common forward, CNV23598 insertion is missing here, B) Insertion reverse, failed to generate a PCR product, C) Deletion reverse confirming *CASP8* exon12 sequence. The red arrow indicates the location where CNV23598 should appear if present.

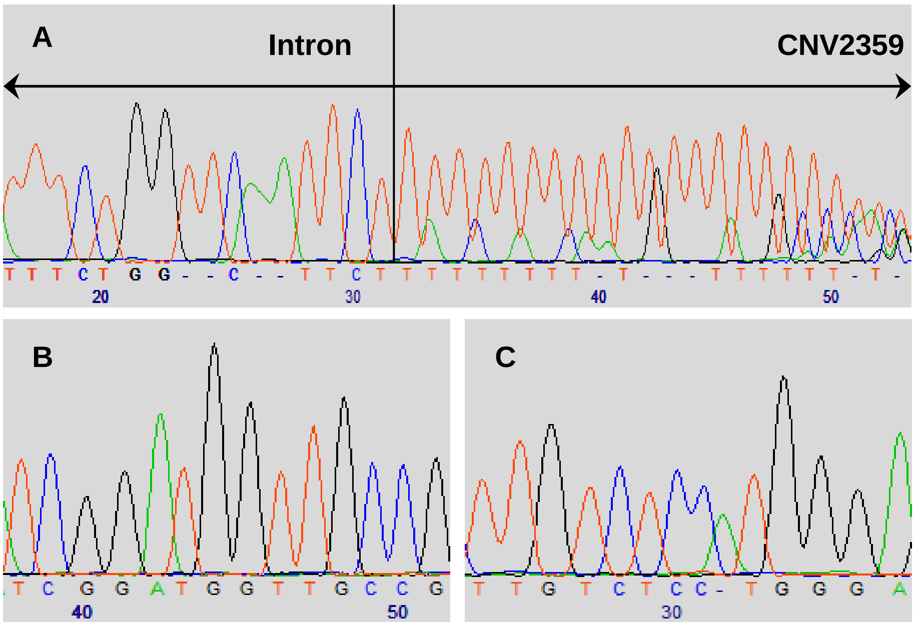


Figure 3.17 Heterozygous CNV23598

DNA sequencing of a sample with heterozygous genotype using the primers designed for the GAP PCR. A) Common forward, the figure indicates the start of CNV23598. The double sequence indicates the presence of a deletion product as well. B) Insertion reverse confirming the CNV23598 sequence, C) Deletion reverse confirming *CASP8* exon12 sequence.

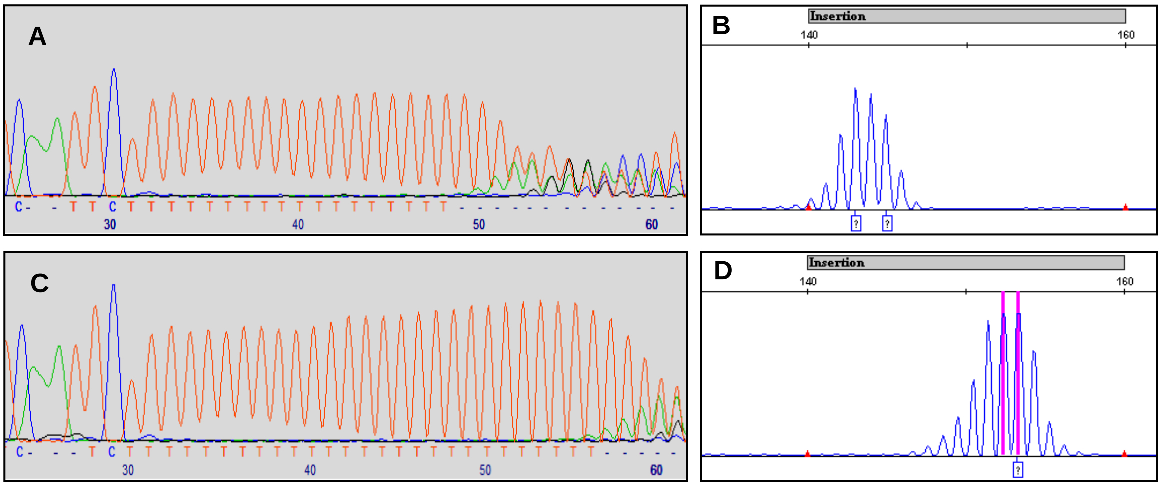


Figure 3.18 CNV23598 22T/32T alleles

DNA sequencing and QF-PCR results of sample C121 showing the 22/~32 Ts at the 3'start of the CNV23598. A & B represent the 22Ts repeat. CD represent ~32Ts repeat.

3.2.3.5 A SNP in the insertion sequence suggests the presence of an apparent extra insertion allele

An interesting event was observed while analysing the results of the heterozygous samples. A single heterozygous SNP c.898+5147A/G was detected in 5 of the heterozygous samples in the insertion sequence. The SNP was detected through the analysis of the sequences produced by the insertion reverse primer. The common forward sequence cannot be analysed downstream of the stretch of Ts at the beginning of the CNV23598 sequence as the DNA sequence becomes noisy due to Taq polymerase slippage. In order to confirm the SNP, an internal forward primer was designed (identical to the fluorescent insertion reverse primer, but in the forward direction) downstream of the stretch of Ts (Figure 3.13). The SNP was then confirmed by bi-directional sequencing (Figure 3.19). The presence of this SNP was remarkable because the heterozygous samples should in theory have one insertion allele and one deletion allele.

The apparent extra insertion copy was also confirmed by analysing the QF-PCR results. Two of the heterozygous samples have shown the presence of the 2 alleles of the T stretch 22/32 in addition to the deletion allele (Figure 3.20). It is worth noting here that both of these samples had the heterozygous c.898+5147A/G SNP. However, the A/G SNP and the 32T allele did not associate together as 2 of the samples carrying the SNP had he 22T insertion allele only. These apparent extra insertion copies could account for the genotyped population being inconsistent with HWE by having too many heterozygous genotypes (Section 3.2.3.4). Table 3.12 summarises the genotyping results of the 16 samples using GAP-PCR, fluorescent GAP PCR and sequencing. The table also shows the samples where 3 copies of CNV23598 were predicted.

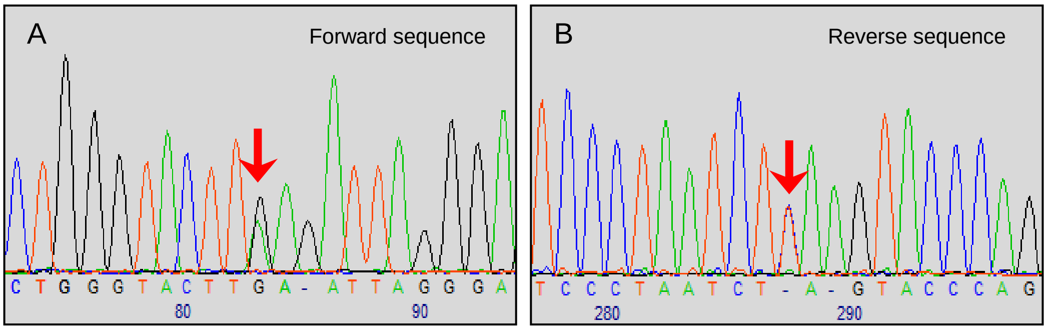


Figure 3.19 **Novel c.898+5147A/G identified in CNV23598 insertion sequence**

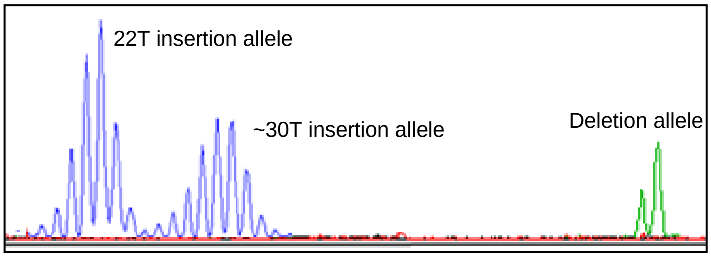


Figure 3.20 **Three copies of CNV23598 using QF-PCR**

Table 3.12 **GAP-PCR, QFPCR and sequencing for CNV23598 in 16 samples**

	Genotype		
Sample	**GAP PCR**	**QFPCR***	**Sequencing****
C004	Ins/Ins	Ins/Ins	Ins/Ins
C009	Ins/Ins	Ins/Ins	Ins/Ins
C010	Ins/Del	Ins/Del	**Ins/Ins/del**
C017	Ins/Ins	Ins/Ins	Ins/Ins
C021	Ins/Ins	Ins/Ins	Ins/Ins
C047	Ins/Ins	Ins/Ins	Ins/Ins
C066	Ins/Del	**Ins/Ins/del**	**Ins/Ins/del**
C069	Ins/Del	Ins/Del	**Ins/Ins/del**
C091	Ins/Ins	Ins/Ins	Ins/Ins
C099	Ins/Del	Ins/Del	Ins/Del
C106	Del/Del	Del/Del	Del/Del
C111	Ins/Del	Ins/Del	Ins/Del
C115	Ins/Del	Ins/Del	Ins/Del
C117	Ins/Del	**Ins/Ins/del**	**Ins/Ins/del**
C119	Del/Del	Del/Del	Del/Del
C121	Ins/Del	**Ins/Ins/del**	**Ins/Ins/del**

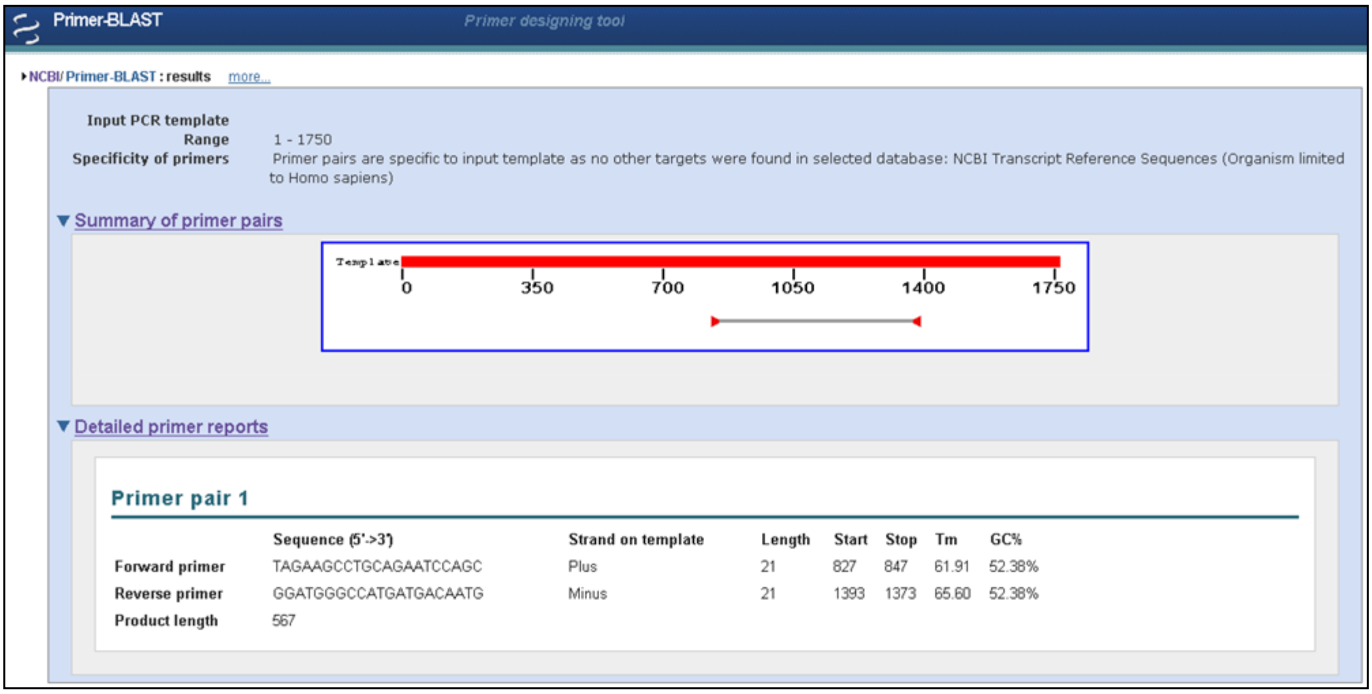
* Ins/Ins/Del genotype with QF-PCR was based on having 2 insertion peaks of different size in addition to the deletion peak

** Ins/Ins/del genotype with sequencing was based on having the c.898+5147A/G SNP in a heterozygous sample

3.2.3.6 *In silico* analysis

The CNV23598 sequence was found to be common across the genome with high homolog, thus, it was hypothesised that the apparent observation of 3 alleles might be due to non-specific amplification.. Several *in silico* approaches were taken to investigate the possibility of non-specific amplification.To begin with, the specificity of the GAP PCR primers (both gel-based and fluorescent) were initially checked by primer BLAST as described in Section 2.2.5.1 (Figure 3.21 and Figure 3.22). This predicted high specificity of the primers to their target sequence, suggesting that non-specific PCR products would not be expected. Following that, the sequence upstream of CNV23598, up to the beginning of the common forward primer sequence, was checked by nucleotide BLAST (http://blast.ncbi.nlm.nih.gov/Blast.cgi) and BLAT (http://genome.ucsc.edu/cgi-bin/hgBlat?command=start) in order to confirm the specificity of the area where the common forward primer was designed. The results showed that this sequence is highly specific to *CASP8* intron 11 (Figure 3.23).

Next, the region amplified using the common forward primer and the insertion reverse primer was checked by BLAST and BLAT, which resulted in several regions with very high homology (Figure 3.24). The high homology areas were shared only with CNV23598 sequence. Moreover, analysing the 15 highest hits in detail revealed that none of the sequences could explain the observation of the single A/G SNP or the extended stretch of Ts (~30) in our sequence. It is worth mentioning that chr4:15793565-15794093(-) had the same A/G SNP, however, it had several more variants in addition. Moreover, chr12:90260871-90261372(+) had the extended T stretch (30Ts) but with 2 more SNPs. Additionally, none matched the upstream sequence in *CASP8* intron 11, and so would not be predicted to amplify with these primers.



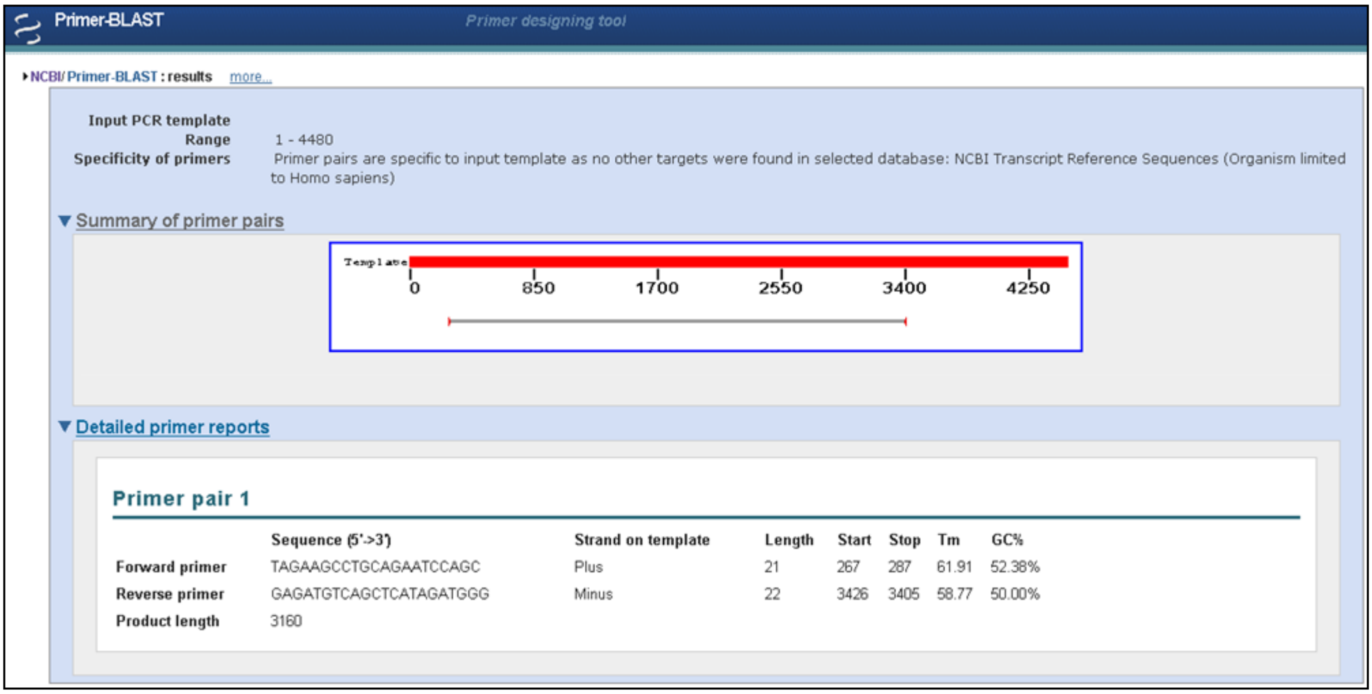
Input PCR template

Range 1 - 1750

Specificity of primers Primer pairs are specific to input template as no other targets were found in selected database: NCBI Transcript Reference Sequences (Organism limited to Homo sapiens)

Primer pair 1

	Sequence (5'->3')	Strand on template	Length	Start	Stop	Tm	GC%
Forward primer	TAGAAGCCTGCAGAATCCAGC	Plus	21	827	847	61.91	52.38%
Reverse primer	GGATGGGCCATGATGACAATG	Minus	21	1393	1373	65.60	52.38%
Product length	567						



Input PCR template

Range 1 - 4480

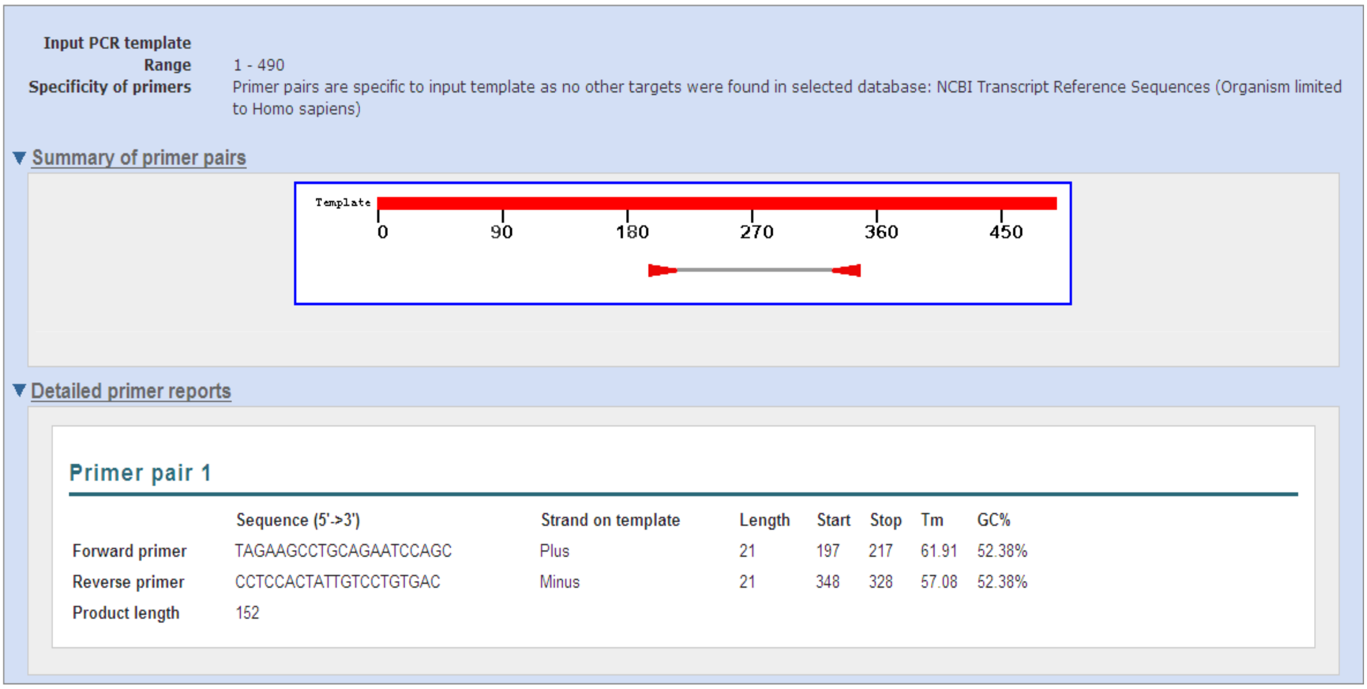
Specificity of primers Primer pairs are specific to input template as no other targets were found in selected database: NCBI Transcript Reference Sequences (Organism limited to Homo sapiens)

Primer pair 1

	Sequence (5'->3')	Strand on template	Length	Start	Stop	Tm	GC%
Forward primer	TAGAAGCCTGCAGAATCCAGC	Plus	21	267	287	61.91	52.38%
Reverse primer	GAGATGTCAGCTCATAGATGGG	Minus	22	3426	3405	58.77	50.00%
Product length	3160						

Figure 3.21 Primer BLAST for CNV23598 gel-based GAP PCR primers

The results showed high predicted specificity of the primers to their target sequence. Note that the amplicon size of the common forward with the deletion reverse includes CNV23598 sequence.



Input PCR template

Range 1 - 490

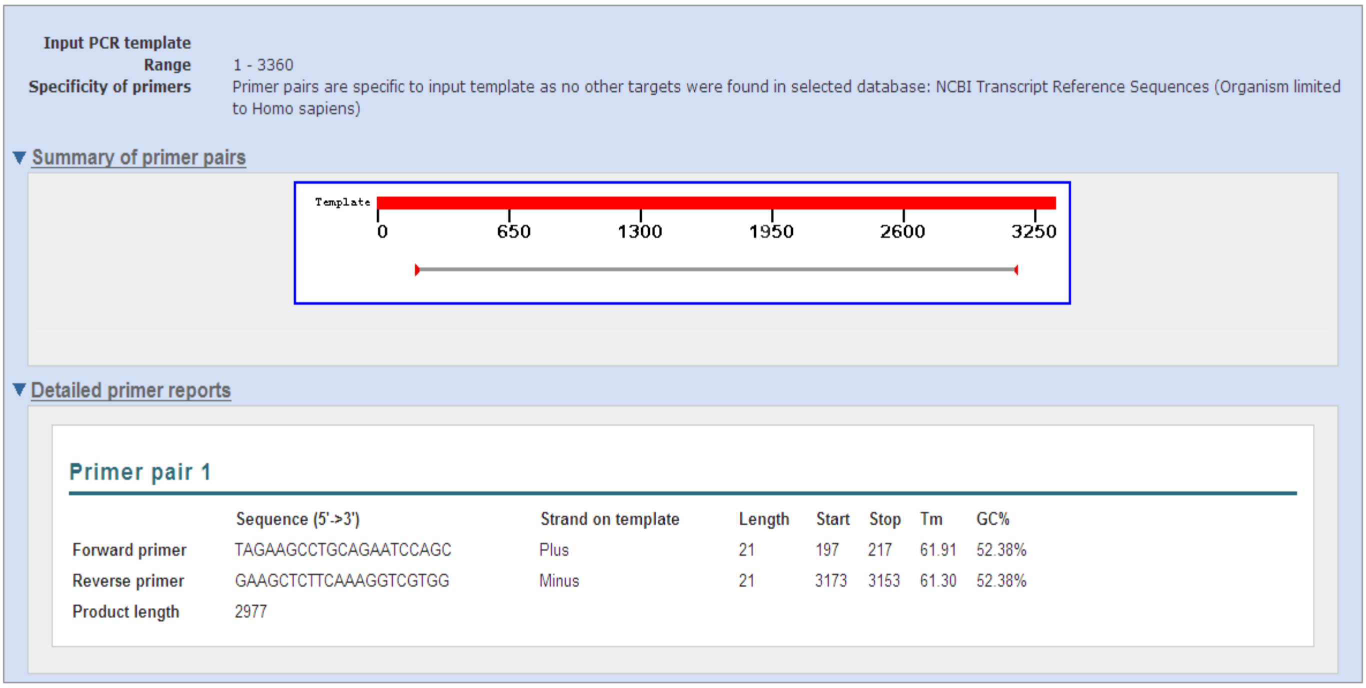
Specificity of primers Primer pairs are specific to input template as no other targets were found in selected database: NCBI Transcript Reference Sequences (Organism limited to Homo sapiens)

▼ Summary of primer pairs

▼ Detailed primer reports

Primer pair 1

	Sequence (5'->3')	Strand on template	Length	Start	Stop	Tm	GC%
Forward primer	TAGAAGCCTGCAGAATCCAGC	Plus	21	197	217	61.91	52.38%
Reverse primer	CCTCCACTATTGTCCTGTGAC	Minus	21	348	328	57.08	52.38%
Product length	152						



Input PCR template

Range 1 - 3360

Specificity of primers Primer pairs are specific to input template as no other targets were found in selected database: NCBI Transcript Reference Sequences (Organism limited to Homo sapiens)

▼ Summary of primer pairs

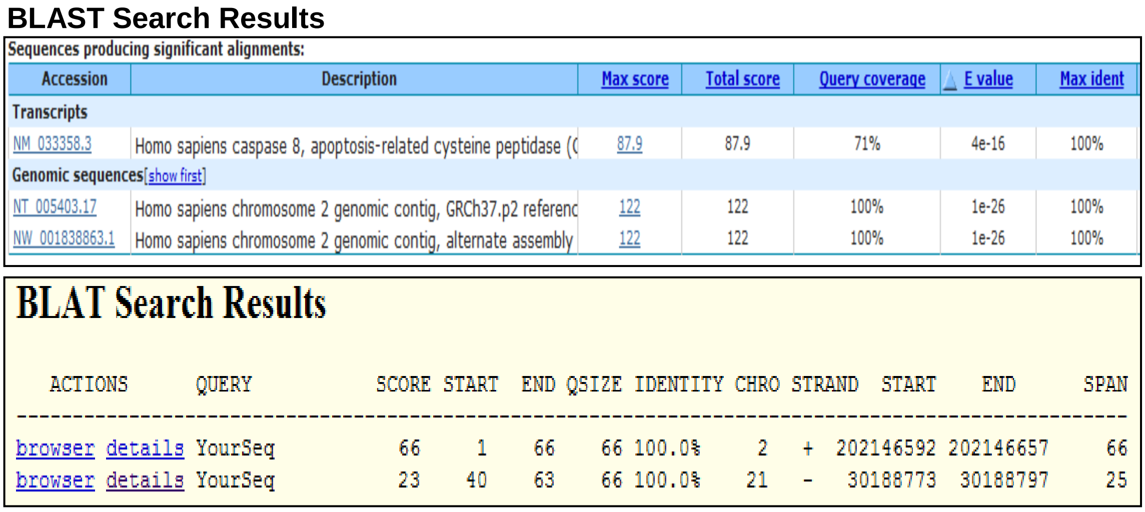
▼ Detailed primer reports

Primer pair 1

	Sequence (5'->3')	Strand on template	Length	Start	Stop	Tm	GC%
Forward primer	TAGAAGCCTGCAGAATCCAGC	Plus	21	197	217	61.91	52.38%
Reverse primer	GAAGCTCTTCAAAGGTCGTGG	Minus	21	3173	3153	61.30	52.38%
Product length	2977						

Figure 3.22 Primer BLAST for CNV23598 fluorescent GAP PCR primers

The results showed high predicted specificity of the primers to their target sequence. Note that the amplicon size of the common forward with the deletion reverse includes CNV23598 sequence.



BLAST Search Results

Sequences producing significant alignments:

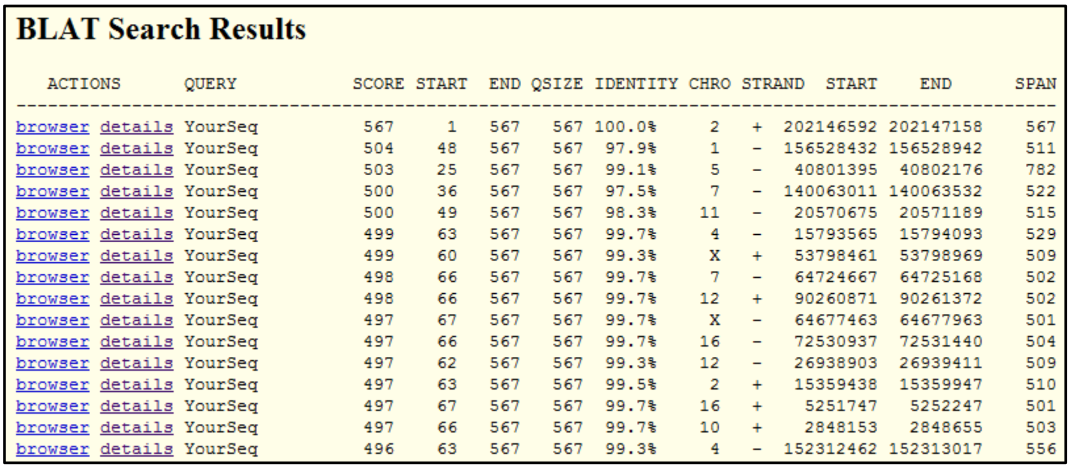
Accession	Description	Max score	Total score	Query coverage	E value	Max ident
Transcripts						
NM_033358.3	Homo sapiens caspase 8, apoptosis-related cysteine peptidase ((	87.9	87.9	71%	4e-16	100%
Genomic sequences[show first]						
NT_005403.17	Homo sapiens chromosome 2 genomic contig, GRCh37.p2 referenc	122	122	100%	1e-26	100%
NW_001838863.1	Homo sapiens chromosome 2 genomic contig, alternate assembly	122	122	100%	1e-26	100%

BLAT Search Results

```
   ACTIONS      QUERY           SCORE START  END QSIZE IDENTITY CHRO STRAND  START    END      SPAN
---------------------------------------------------------------------------------------------------
browser details YourSeq            66     1    66    66 100.0%     2   +   202146592 202146657    66
browser details YourSeq            23    40    63    66 100.0%    21   -    30188773  30188797    25
```

Figure 3.23 BLAST and BLAT of the 66 nucleotides upstream CNV23598

The results confirmed the high specificity of the 66 nucleotides (upstream CNV23598 in *CASP8* intron11) to the target sequence.



BLAT Search Results

```
   ACTIONS      QUERY           SCORE START  END QSIZE IDENTITY CHRO STRAND  START    END      SPAN
---------------------------------------------------------------------------------------------------
browser details YourSeq           567     1   567   567 100.0%     2   +   202146592 202147158   567
browser details YourSeq           504    48   567   567  97.9%     1   -   156528432 156528942   511
browser details YourSeq           503    25   567   567  99.1%     5   -    40801395  40802176   782
browser details YourSeq           500    36   567   567  97.5%     7   -   140063011 140063532   522
browser details YourSeq           500    49   567   567  98.3%    11   -    20570675  20571189   515
browser details YourSeq           499    63   567   567  99.7%     4   -    15793565  15794093   529
browser details YourSeq           499    60   567   567  99.3%     X   +    53798461  53798969   509
browser details YourSeq           498    66   567   567  99.7%     7   -    64724667  64725168   502
browser details YourSeq           498    66   567   567  99.7%    12   +    90260871  90261372   502
browser details YourSeq           497    67   567   567  99.7%     X   -    64677463  64677963   501
browser details YourSeq           497    66   567   567  99.7%    16   -    72530937  72531440   504
browser details YourSeq           497    62   567   567  99.3%    12   -    26938903  26939411   509
browser details YourSeq           497    63   567   567  99.5%     2   +    15359438  15359947   510
browser details YourSeq           497    67   567   567  99.7%    16   +     5251747   5252247   501
browser details YourSeq           497    66   567   567  99.7%    10   +     2848153   2848655   503
browser details YourSeq           496    63   567   567  99.3%     4   -   152312462 152313017   556
```

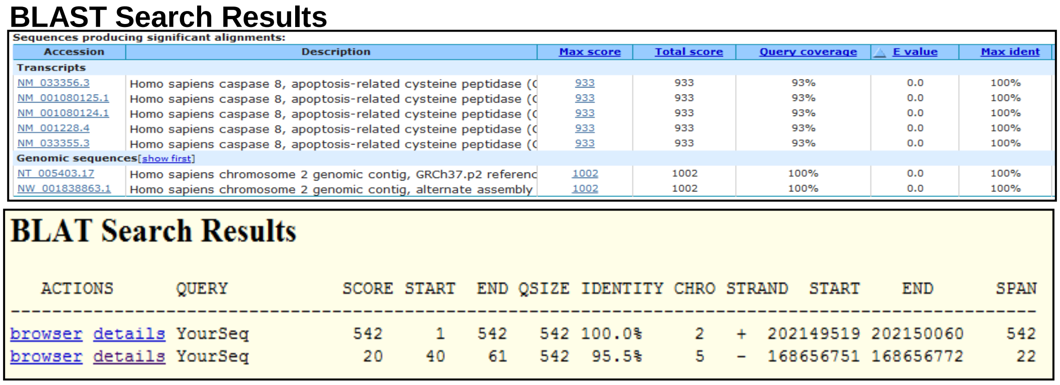
Figure 3.24 BLAT for CNV23598 amplicon

The highest 15 hits from “BLATing” the whole region amplified by common forward and insertion reverse primers.

3.2.3.7 Testing a 3’ GAP PCR assay

Because of the difficulty of designing more primers in the 5’ intron 11 area upstream of CNV23598, it was decided to try and genotype CNV23598 from the 3’ end. This would allow the use of the reverse primer in *CASP8* exon12 as the common primer. *CASP8* exon12 was checked by BLAT and BLAST and results confirmed that this area is unique to *CASP8* (Figure 3.25). The 3’ deletion reverse primer was designed in the same location as the 5’ common forward primer and the 3’ common reverse primer was designed in the same location of the 5’deletion reverse primer (Figure 3.26). Designing a 3’ forward insertion primer was not easy, due to potential non-specific products identified in primer BLAST. The primer with the lowest number of potential non-specific products and the maximum number of mismatches with these products was chosen. It had three potentially unintended targets with sizes significantly different from the intended target (366bp, 2343bp and 3062bp) (Figure 3.27). The unintended targets were the results of the forward primer serving both as forward and reverse. Aligning the sequences of the potentially unintended products with the *CASP8* target region using the specialised BLAST “align” (http://blast.ncbi.nlm.nih.gov/Blast.cgi) has shown that the unintended target regions were completely different from the actual target as no significant similarity was found.

The primers were then used to amplify samples that were shown to be heterozygous, homozygous insertion and homozygous deletion (Figure 3.28). The results confirmed the genotypes of the samples as identified previously; however, the size of the insertion product was consistently larger than the expected product by ~450bp. As shown in Figure 3.26, the expected insertion product was 406bp but the results showed ~850bp insertion products (Figure 3.28). The results were confirmed using different samples and the PCR products were then sequenced to verify their specificity. DNA sequencing results confirmed the specificity of the PCR products to the target area, however, two main repeat sequences at the 3’end of the CNV23598 were found to be expanded (Figure 3.29). This expansion of repeats could account for the larger PCR products obtained with the 3’ forward insertion and the 3’ common reverse primers.



BLAST Search Results

Sequences producing significant alignments:

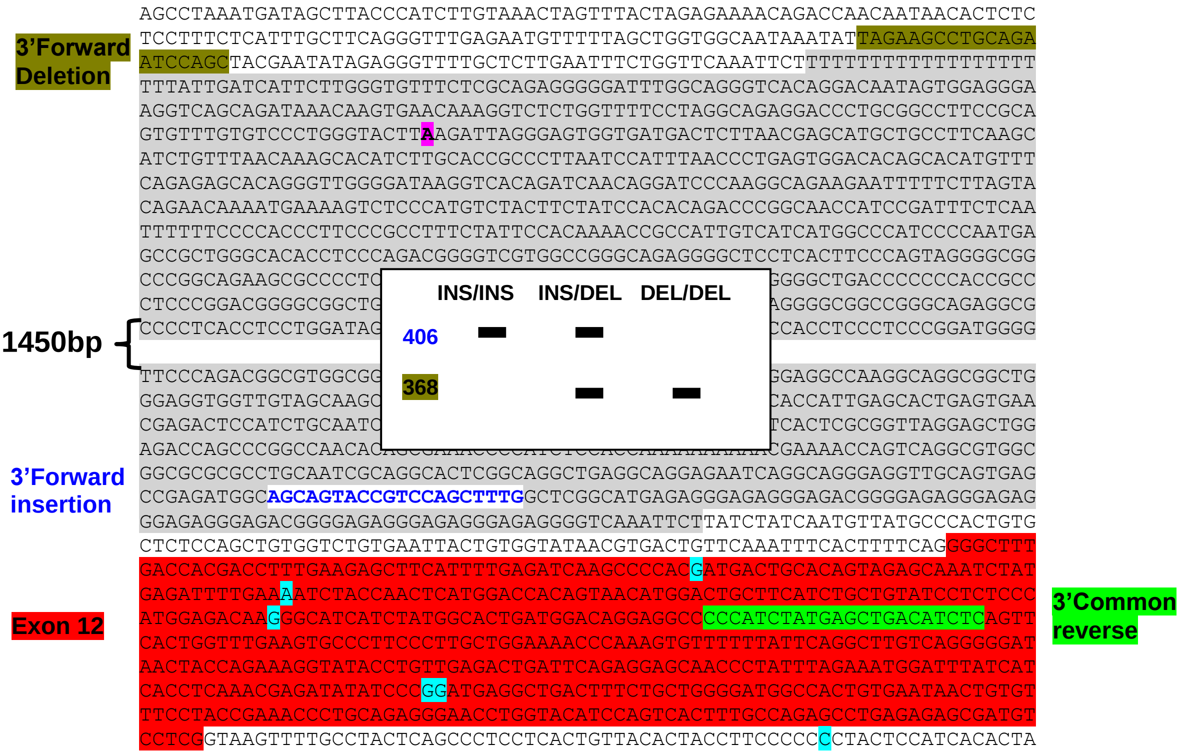
Accession	Description	Max score	Total score	Query coverage	E value	Max ident
Transcripts						
NM_033356.3	Homo sapiens caspase 8, apoptosis-related cysteine peptidase (C	933	933	93%	0.0	100%
NM_001080125.1	Homo sapiens caspase 8, apoptosis-related cysteine peptidase (C	933	933	93%	0.0	100%
NM_001080124.1	Homo sapiens caspase 8, apoptosis-related cysteine peptidase (C	933	933	93%	0.0	100%
NM_001228.4	Homo sapiens caspase 8, apoptosis-related cysteine peptidase (C	933	933	93%	0.0	100%
NM_033355.3	Homo sapiens caspase 8, apoptosis-related cysteine peptidase (C	933	933	93%	0.0	100%
Genomic sequences[show first]						
NT_005403.17	Homo sapiens chromosome 2 genomic contig, GRCh37.p2 referenc	1002	1002	100%	0.0	100%
NW_001838863.1	Homo sapiens chromosome 2 genomic contig, alternate assembly	1002	1002	100%	0.0	100%

BLAT Search Results

```
   ACTIONS      QUERY            SCORE START   END QSIZE IDENTITY CHRO STRAND  START     END       SPAN
---------------------------------------------------------------------------------------------------
browser details YourSeq            542     1   542   542 100.0%     2   +   202149519 202150060   542
browser details YourSeq             20    40    61   542  95.5%     5   -   168656751 168656772    22
```

Figure 3.25 BLAST and BLAT for *CASP8* exon 12

The results of BLAST and BLAT confirmed the uniqueness of *CASP8* exon 12 sequence

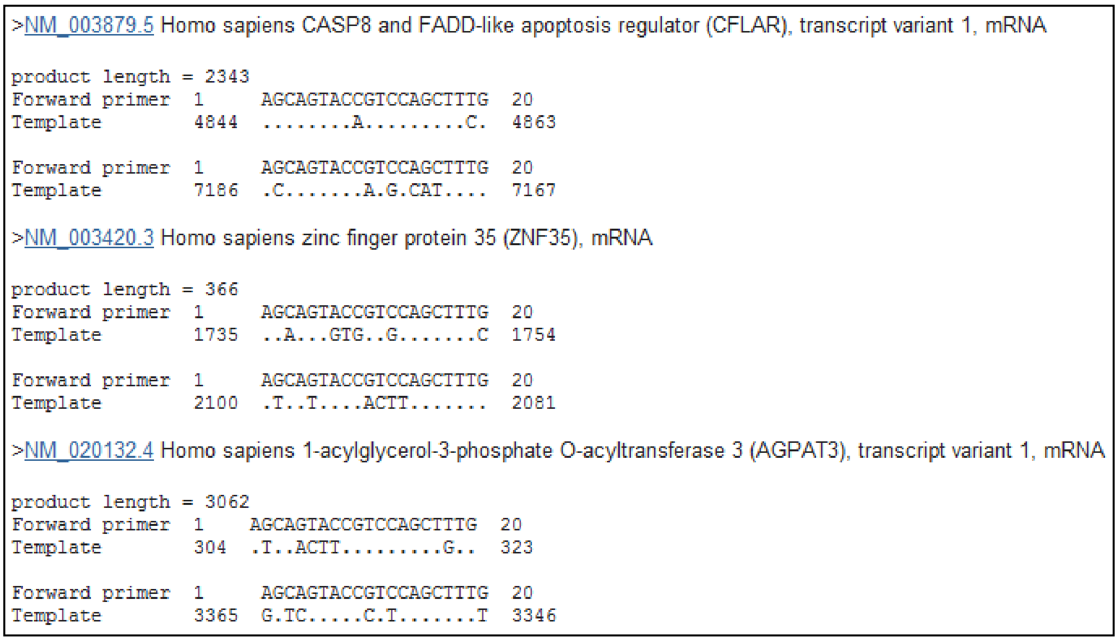


```
AGCCTAAATGATAGCTTACCCATCTTGTAAACTAGTTTACTAGAGAAAACAGACCAACAATAACACTCTC
TCCTTTCTCATTTGCTTCAGGGTTTGAGAATGTTTTTAGCTGGTGGCAATAAATATTAGAAGCCTGCAGA
ATCCAGCTACGAATATAGAGGGTTTTGCTCTTGAATTTCTGGTTCAAATTCTTTTTTTTTTTTTTTTTT
TTTATTGATCATTCTTGGGTGTTTCTCGCAGAGGGGGATTTGGCAGGGTCACAGGACAATAGTGGAGGGA
AGGTCAGCAGATAAACAAGTGAACAAAGGTCTCTGGTTTTCCTAGGCAGAGGACCCTGCGGCCTTCCGCA
GTGTTTGTGTCCCTGGGTACTTAAGATTAGGGAGTGGTGATGACTCTTAACGAGCATGCTGCCTTCAAGC
ATCTGTTTAACAAAGCACATCTTGCACCGCCCTTAATCCATTTAACCCTGAGTGGACACAGCACATGTTT
CAGAGAGCACAGGGTTGGGGATAAGGTCACAGATCAACAGGATCCCAAGGCAGAAGAATTTTTCTTAGTA
CAGAACAAAATGAAAAGTCTCCCATGTCTACTTCTATCCACACAGACCCGGCAACCATCCGATTTCTCAA
TTTTTTCCCCACCCTTCCCGCCTTTCTATTCCACAAAACCGCCATTGTCATCATGGCCCATCCCCAATGA
GCCGCTGGGCACACCTCCCAGACGGGGTCGTGGCCGGGCAGAGGGGCTCCTCACTTCCCAGTAGGGGCGG
CCCGGCAGAAGCGCCCCTC                              GGGGCTGACCCCCCCACCGCC
CTCCCGGACGGGGCGGCTG                              AGGGGCGGCCGGGCAGAGGCG
CCCCTCACCTCCTGGATAG                              CCACCTCCCTCCCGGATGGGG

TTCCCAGACGGCGTGGCGG                              GGAGGCCAAGGCAGGCGGCTG
GGAGGTGGTTGTAGCAAGC                              CACCATTGAGCACTGAGTGAA
CGAGACTCCATCTGCAATC                              TCACTCGCGGTTAGGAGCTGG
AGACCAGCCCGGCCAACACAGCCGAAACCCCATCTCCACCAAAAAAAAAACGAAAACCAGTCAGGCGTGGC
GGCGCGCGCCTGCAATCGCAGGCACTCGGCAGGCTGAGGCAGGAGAATCAGGCAGGGAGGTTGCAGTGAG
CCGAGATGGCAGCAGTACCGTCCAGCTTTGGCTCGGCATGAGAGGGAGAGGGAGACGGGGAGAGGGAGAG
GGAGAGGGAGACGGGGAGAGGGAGAGGGAGAGGGGTCAAATTCTTATCTATCAATGTTATGCCCACTGTG
CTCTCCAGCTGTGGTCTGTGAATTACTGTGGTATAACGTGACTGTTCAAATTTCACTTTTCAGGGGCTTT
GACCACGACCTTTGAAGAGCTTCATTTTGAGATCAAGCCCCACGATGACTGCACAGTAGAGCAAATCTAT
GAGATTTTGAAAATCTACCAACTCATGGACCACAGTAACATGGACTGCTTCATCTGCTGTATCCTCTCCC
ATGGAGACAAGGGCATCATCTATGGCACTGATGGACAGGAGGCCCCCATCTATGAGCTGACATCTCAGTT
CACTGGTTTGAAGTGCCCTTCCCTTGCTGGAAAACCCAAAGTGTTTTTTATTCAGGCTTGTCAGGGGGAT
AACTACCAGAAAGGTATACCTGTTGAGACTGATTCAGAGGAGCAACCCTATTTAGAAATGGATTTATCAT
CACCTCAAACGAGATATATCCCGGATGAGGCTGACTTTCTGCTGGGGATGGCCACTGTGAATAACTGTGT
TTCCTACCGAAACCCTGCAGAGGGAACCTGGTACATCCAGTCACTTTGCCAGAGCCTGAGAGAGCGATGT
CCTCGGTAAGTTTTGCCTACTCAGCCCTCCTCACTGTTACACTACCTTCCCCCCCTACTCCATCACACTA
```

Figure 3.26 CNV23598 3’GAP PCR

CNV23598 is highlighted in grey, exon 12 in red and SNPs in turquoise. Primers are highlighted in different colours as indicated. The sizes of the expected products and the genotypes are represented in the box.



```
>NM_003879.5 Homo sapiens CASP8 and FADD-like apoptosis regulator (CFLAR), transcript variant 1, mRNA

product length = 2343
Forward primer  1     AGCAGTACCGTCCAGCTTTG  20
Template        4844  ........A.........C.  4863

Forward primer  1     AGCAGTACCGTCCAGCTTTG  20
Template        7186  .C.......A.G.CAT....  7167

>NM_003420.3 Homo sapiens zinc finger protein 35 (ZNF35), mRNA

product length = 366
Forward primer  1     AGCAGTACCGTCCAGCTTTG  20
Template        1735  ..A...GTG..G.......C  1754

Forward primer  1     AGCAGTACCGTCCAGCTTTG  20
Template        2100  .T..T....ACTT.......  2081

>NM_020132.4 Homo sapiens 1-acylglycerol-3-phosphate O-acyltransferase 3 (AGPAT3), transcript variant 1, mRNA

product length = 3062
Forward primer  1    AGCAGTACCGTCCAGCTTTG  20
Template        304  .T..ACTT.........G..  323

Forward primer  1     AGCAGTACCGTCCAGCTTTG  20
Template        3365  G.TC.....C.T.......T  3346
```

Figure 3.27 Potential unintended products with the 3' insertion forward primer

These non-specific targets are the products of the forward primer only acting in both directions. All the sizes differ significantly from the expected size of the specific target and they have several mismatches with the primer in both directions.

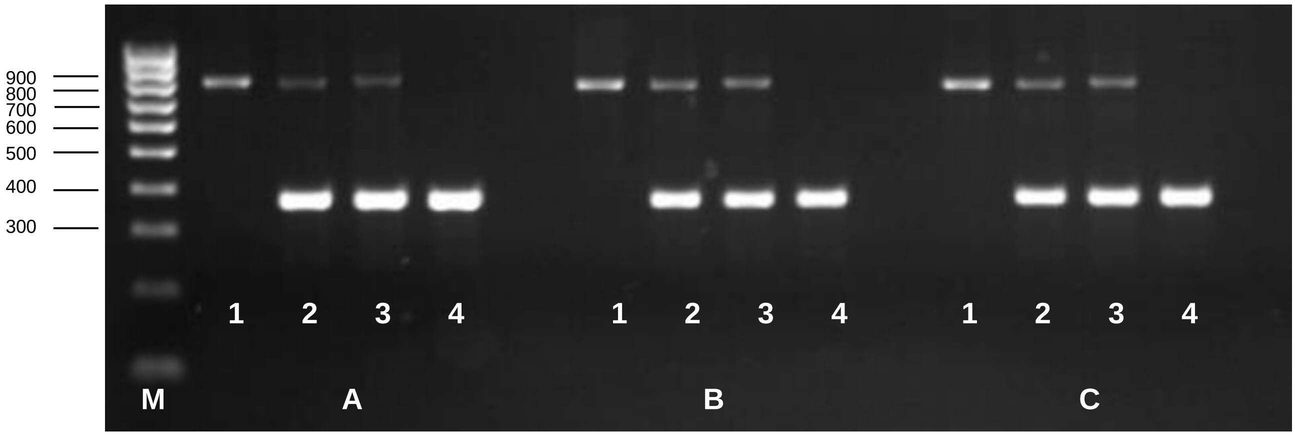
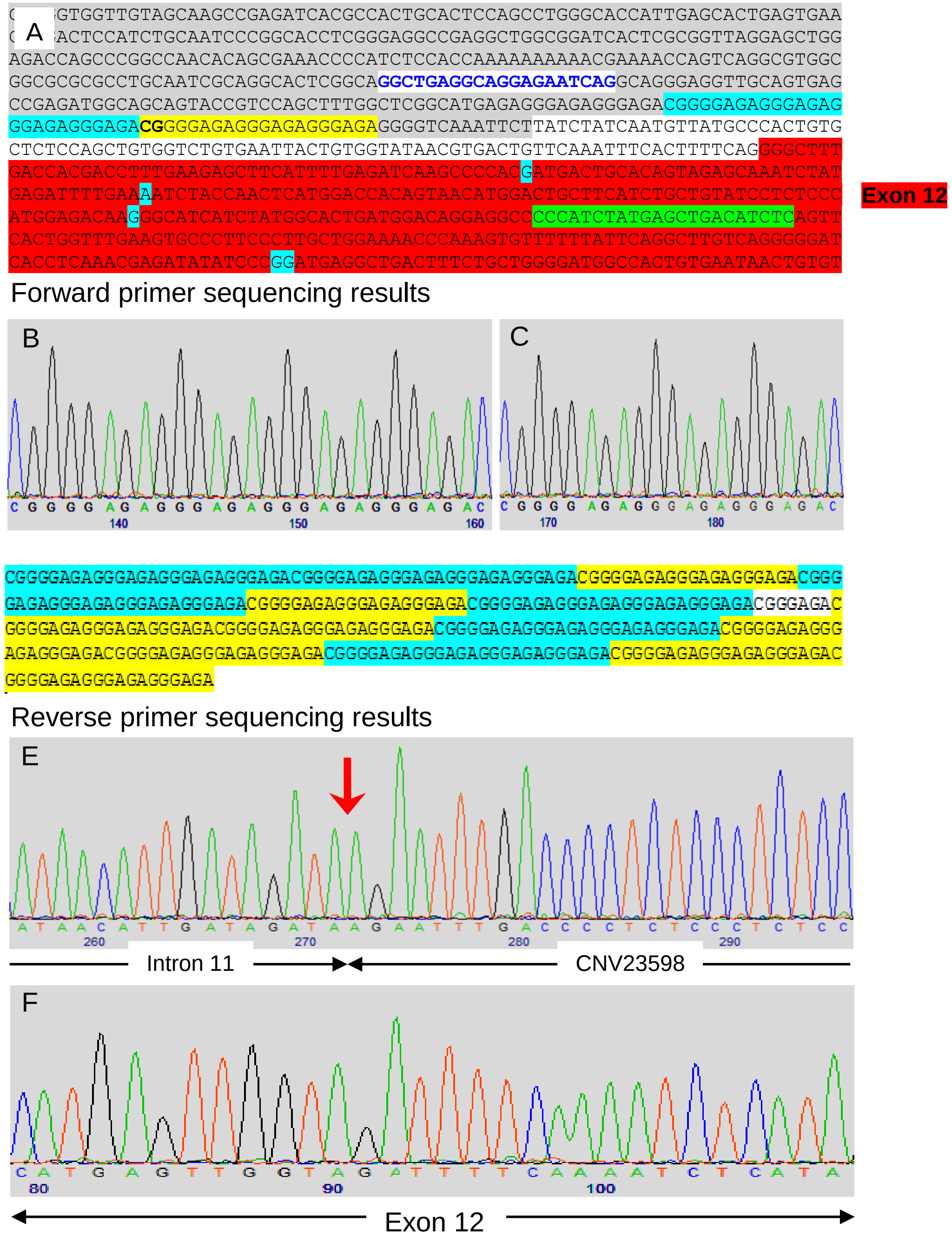


Figure 3.28 CNV23598 3' GAP PCR gel electrophoresis

Genotyping CNV23598 from the 3'end. The genotypes of the samples were confirmed. Different conditions (A, B and C) were used and they all gave specific products. The size of the deletion product was as expected (368bp). The size of the insertion product was consistently larger (~850bp) than expected (406bp). M: Hyperladder IV Samples 1-homozygous insertion, 2&3- heterozygous and 4- homozygous deletion.



```
C GTGGTTGTAGCAAGCCGAGATCACGCCACTGCACTCCAGCCTGGGCACCATTGAGCACTGAGTGAA
C ACTCCATCTGCAATCCCGGCACCTCGGGAGGCCGAGGCTGGCGGATCACTCGCGGTTAGGAGCTGG
AGACCAGCCCGGCCAACACAGCGAAACCCCATCTCCACCAAAAAAAAAACGAAAACCAGTCAGGCGTGGC
GGCGCGCGCCTGCAATCGCAGGCACTCGGCAGGCTGAGGCAGGAGAATCAGGCAGGGAGGTTGCAGTGAG
CCGAGATGGCAGCAGTACCGTCCAGCTTTGGCTCGGCATGAGAGGGAGAGGGAGACGGGGAGAGGGAGAG
GGAGAGGGAGACGGGGAGAGGGAGAGGGAGAGGGGTCAAATTCTTATCTATCAATGTTATGCCCACTGTG
CTCTCCAGCTGTGGTCTGTGAATTACTGTGGTATAACGTGACTGTTCAAATTTCACTTTTCAGGGGCTTT
GACCACGACCTTTGAAGAGCTTCATTTTGAGATCAAGCCCCACGATGACTGCACAGTAGAGCAAATCTAT
GAGATTTTGAAAATCTACCAACTCATGGACCACAGTAACATGGACTGCTTCATCTGCTGTATCCTCTCCC
ATGGAGACAAGGGCATCATCTATGGCACTGATGGACAGGAGGCCCCCATCTATGAGCTGACATCTCAGTT
CACTGGTTTGAAGTGCCCTTCCCTTGCTGGAAAACCCAAAGTGTTTTTTATTCAGGCTTGTCAGGGGGAT
CACCTCAAACGAGATATATCCCGGATGAGGCTGACTTTCTGCTGGGGATGGCCACTGTGAATAACTGTGT
```

```
CGGGGAGAGGGAGAGGGAGAGGGAGACGGGGAGAGGGAGAGGGAGAGGGAGACGGGGAGAGGGAGAGGGAGACGGG
GAGAGGGAGAGGGAGAGGGAGACGGGGAGAGGGAGAGGGAGACGGGGAGAGGGAGAGGGAGAGGGAGACGGGAGAC
GGGGAGAGGGAGAGGGAGACGGGGAGAGGGAGAGGGAGACGGGGAGAGGGAGAGGGAGAGGGAGACGGGGAGAGGG
AGAGGGAGACGGGGAGAGGGAGAGGGAGACGGGGAGAGGGAGAGGGAGAGGGAGACGGGGAGAGGGAGAGGGAGAC
GGGGAGAGGGAGAGGGAGA
```

Figure 3.29 Repeat sequences at 5' end of CNV23598

Sequencing results from the 3' forward insertion & 3' common reverse primers showing an expansion of two main repeat sequences B &C. A) The target area with the two repeat sequences B & C highlighted in turquoise and yellow respectively D) The sequence of the expansion that was analysable by sequencing and it contains repeats of B&C sequences; BBCBCBCGGGAGACCBCCBCC. E and F represent the sequencing results of the 3' common reverse primer. The sequences confirm the specificity of the PCR products.

3.2.3.8 Long range PCR *CASP8* exon11- exon12

The size of the region of *CASP8* between exon11 and 12 (including CNV23598) is 8491bp. In order to try and solve the problems of genotyping CNV23598, we decided to amplify the whole region by long range PCR and then sequence with internal primers. Phusion Taq polymerase have been successfully used in our lab to amplify PCR products up to 10Kb in size, however, several attempts using different DNA samples, primers and PCR conditions failed completely to give any specific products.

3.3 Discussion

As mentioned earlier, our lab had some preliminary data indicating an association between *CASP8* SNPs and CRC risk. In order to further investigate the potential role of inherited *CASP8* variants in affecting CRC risk, the following was performed; firstly, rs3834129 (an ins/del variant in the promoter region of *CASP8* previously implicated in CRC risk) (Sun *et al.* 2007) was genotyped in 1193 CRC cases and 1388 controls from Sheffield, Leeds, Dundee and Utah. Secondly, the *CASP8* coding region, in addition to intron/exon boundaries, the promoter region and 3'UTR were sequenced in 94 CRC cases from the Sheffield cohort.

The SNP rs3834129 was genotyped using TaqMan assays. Several QC assessments were performed to validate the TaqMan assay and the genotyping data. In summary, the QC measures indicated a reasonable quality of the genotyping results. Chi-square association test was then performed and the results have shown that there was no significant difference in genotype frequencies between cases and controls. These results contradict the previous results from the Chinese population (Sun *et al.* 2007) which showed an association between rs3834129 and multiple cancer risks including CRC. However, the results are in concordance with the published data on the European and multi-ethnic American populations which found no effect of rs3834129 on CRC risk (Haiman *et al.* 2008, Pittman *et al.* 2008). Based on the size of the analysed cohort, we had ~98% power (α=0.05) to detect an effect of similar size (OR=0.75/ copy of deletion allele) as described by Sun and colleagues. Possible explanations for the discrepancy between our results (and the European and the American studies) in comparison to the original study in the Chinese population could be due to different frequencies and effects of genetic and environmental modifier factors between the different populations, or due to the initial association being a false positive. Despite the lack of association overall, rs3834129 genotyping results are currently being used in a more in depth analysis of *CASP8* inherited variants and CRC risk. This further analysis is being performed both to fine map associations and to investigate various subgroups of CRC cases (Curtin *et al.* manuscript in preparation).

CASP8 sequencing was performed to try and identify rare coding variants that might affect CRC risk. Association studies are usually performed using a “tagging SNP” approach which depends on common and known SNPs. Therefore, rare or unknown SNPs could have been easily missed. Sequencing 94 cases yields an 85% probability of identifying variants with a minor allele frequency (MAF) of 0.01. Normal DNA extracted from the peripheral blood of 94 CRC cases from the Sheffield population was sequenced and six novel variants were identified. The first variant (-565 G>A) was within the promoter region of *CASP8*, however, it did not affect any of the known functional sequences identified in that area. Moreover, *in silico* analysis did not predict the introduction of any transcription factors binding sites or the presence of any regulatory regions. The genomic location was not conserved across 5 mammalian species. Therefore, it was not predicted to have an effect on *CASP8* function. The second variant was identified only in one case and was a 115 bp deletion in exon 3 (c.1-8338Del115). Despite this region having a potential regulatory role, as predicted by the presence of DNase cluster region and histone marks which could reflect transcriptional enhancers and promoters (Crawford *et al.* 2006), the following observations led to the prediction that it was unlikely to be functionally important. Firstly, *CASP8* has 8 different isoforms and exon 3 is only expressed in the G isoform precursor. The deletion results in an in-frame deletion of the first 11 amino acids of isoform G precursor which is not part of the mature protein. Additionally, the deleted 11 amino acids are not part of the known conserved functional domains of CASP8 protein. Furthermore, *in silico* analysis of the 11 amino acids sequence using the database of protein domains, families and functional sites Prosite (http://www.expasy.ch/prosite/, accessed July, 2011) did not predict any highly conserved and functional domains. Finally, isoform G is not one of the predominantly expressed *CASP8* isoforms and it is not detected at the protein level (Scaffidi *et al.* 1997). Therefore, this variant did not seem a suitable candidate for further investigation at this stage. The four other variants were intronic, and were not predicted to have an effect on *CASP8* mRNA splicing using 5 different online splice sites prediction tools. Apart from the intron 3 variant c.1-7982A>G, none of the variants were predicted to be in a region with a potential regulatory role. However, the variant c.1-7982A>G was in the same region as the exon 3 115 bp deletion and it was not predicted to be functionally important for the same reasons discussed above.

CNVs were recently shown to play an important role in several human inherited diseases and cancer (Beckmann *et al.* 2007). Moreover, they have been used successfully to predict cancer related genes (Venkatachalam *et al.* 2010). Although the above results do not suggest a role for *CASP8* inherited variants in CRC risk, we decided to investigate the role of CNV23598, reported on the database of genomic variants, because of its physical proximity to the known breast cancer D302H risk variant and due to its unknown frequency.

The genotyping results of 47 selected breast cancer cases and controls suggested a strong association between the insertion allele and the *CASP8* risk haplotype.

However, when 288 CRC cases and controls were investigated, the genotype frequencies differed markedly from HWE with excess heterozygotes. Several molecular techniques (Gel-based and fluorescent GAP PCR, in addition to DNA sequencing) were used in an attempt to validate the genotyping results. The combined results showed some differences from the published reference sequence, and suggested the presence of multiple insertion copies of CNV23598. DNA sequencing results and the use of *in silico* tools indicated that the genotyping assays were specific to the target region. Multi-allelic CNVs are known to occur in the human genome, which could explain the presence of the 3rd copy (McCarroll and Altshuler 2007). Moreover, reported CNVs in the genomic databases usually lack specificity at their breakpoints. Several more CNVs were recently reported on the database of genomic variants which overlap CNV23598, but with variable breakpoints at both ends. This could explain some of the difficulties encountered in the genotyping experiments. Thus, our attempts to accurately genotype CNV23598 using conventional molecular techniques have so far failed. The 47 breast cancer samples described in this project were recently sequenced using a 2nd generation sequencing platform as part of a different project. The analysis of these data may reveal the exact sequence of the CNV flanking regions and resolve some of the inconclusive observations for the data.

4. Colorectal Cancer Molecular Classification

4.1 Introduction

As described in section 1.3.5, recent studies on CRC have shown that the disease is highly heterogeneous from a molecular point of view and should no longer be considered as one disease (Cheng *et al.* 2008, Derks *et al.* 2008, Issa 2008, Ogino and Goel 2008). There are at least three different subtypes that develop through different pathways of genetic and epigenetic instabilities (MSI, CIN and CIMP) and mutations in key cancer genes such as *APC*, *TP53*, *KRAS*, *BRAF* and *PIK3CA* (Figure 1.5). MSI, CIN and CIMP are not entirely mutually exclusive and associations between these genomic instabilities and mutations in the cancer driver genes define the different CRC molecular subtypes.

4.1.1 Associations of molecular events and CRC subtypes

MSI-H was previously shown to be associated with CIMP-H (including *MLH1* methylation) (Weisenberger *et al.* 2006). MSI-H and CIMP-H tumours were both shown to be associated with mutations in *BRAF* (V600E) and *PIK3CA* (exons 9 and 20) (Hawkins *et al.* 2002, Kambara *et al.* 2004, Weisenberger *et al.* 2006, Shen *et al.* 2007, Abubaker *et al.* 2008, Nosho *et al.* 2008). However, the association between MSI-H and *PIK3CA* mutations is controversial and was suggested to occur as a result of MSI-H association with CIMP-H (Nosho *et al.* 2008). MSI-H was also shown to be inversely correlated with CIN and *KRAS* and *TP53* mutations (Samowitz *et al.* 2001b, Ogino and Goel 2008). In general, CIMP tumours (whether H or L, MSI or MSS) were found to associate with *BRAF/KRAS* mutant status (Samowitz *et al.* 2005, Ogino *et al.* 2006b, Weisenberger *et al.* 2006). CIMP-H was found to be inversely correlated with *TP53* mutations and CIN (or significantly associated with low levels of CIN) (Hawkins *et al.* 2002, Goel *et al.* 2007, Cheng *et al.* 2008). CIN was found to be associated with *TP53* mutations (Chang *et al.* 2006). Mutations in

KRAS and *BRAF* oncogenes are mainly mutually exclusive as they both play roles in activating the proliferative MAP Kinase (MAPK) signalling pathway. However, although rare, concomitant *KRAS/BRAF* mutations were shown to have a synergistic effect and were associated with advanced CRC (Oliveira *et al.* 2007).

Based on these molecular associations, several CRC molecular classification systems were proposed with various molecular subtypes (Jass 2007, Issa 2008, Ogino and Goel 2008). The 3 main subtypes include the MSI-H/CIMP-H pathway which accounts for 10-20% of sporadic CRC cases (Chang *et al.* 2006, Weisenberger *et al.* 2006, Shen *et al.* 2007), the CIN/MSS pathway which accounts for 50-70% of sporadic CRC cases (Eshleman *et al.* 1998, Chang *et al.* 2006, Issa 2008) and the CIMP-L/MSS/chromosomally stable pathway which accounts for 10-30% of sporadic CRC cases (Shen *et al.* 2007, Issa 2008, Ogino and Goel 2008). In addition to their distinct molecular characteristics, these CRC molecular subtypes are known to associate with different clinical and pathological features. (Chang *et al.* 2006, Ogino *et al.* 2006b, Shen *et al.* 2007, Issa 2008, Ogino *et al.* 2009). Molecular classification has potential applications in both the clinic and in research. The main goal of molecular classification is to discover molecular biomarkers that can be used to predict prognosis and survival, or to predict treatment response and efficacy and help in the development of targeted cancer therapies (Issa 2008, Ogino and Goel 2008). Moreover, classifying CRC into different molecular subtypes can help in designing more structured studies. For example, studies that have identified associations between certain polymorphisms and specific subtypes of CRC could be missed if CRC was considered a single disorder (Karpinski *et al.* 2010, Slattery *et al.* 2011, Whiffin *et al.* 2011).

At the beginning of this project, several studies had attempted to establish a molecular classification for sporadic CRC (Chang *et al.* 2006, Shen *et al.* 2007, Cheng *et al.* 2008, Derks *et al.* 2008). The studies used different techniques to identify CIN, including conventional and BAC CGH, DNA ploidy analysis by flow cytometry and SNP arrays. Additionally, CIMP was investigated with different panels of CpG island methylation markers and using different technologies, not all of which were quantitative. MSI testing has been more standardised, however, it was mainly performed using the old NCI consensus panel (Section 1.3.3.2.2).

The main limitation of CRC molecular classification studies was the lack of uniformity and standardisation in investigating and defining the molecular characteristics, especially global events such as CIMP and CIN (Issa 2008, Ogino and Goel 2008). This makes it a major challenge to compare results from different studies (Ogino and Goel 2008). The aims of the work described in this chapter were; firstly, to investigate MSI using a panel of 5 mononucleotide microsatellite markers according to the revised NCI criteria on MSI testing (Bacher *et al.* 2004, Umar *et al.* 2004). This panel was shown to be more accurate in MSI testing compared to the old NCI panel (Murphy *et al.* 2006). Secondly, to identify the common mutations in the key genes *APC*, *TP53*, *KRAS*, *BRAF* and *PIK3CA*. Thirdly, to investigate CIMP using a panel of 8 methylation markers that were selected from different CIMP studies and have been validated in a large independent population based study (Ogino *et al.* 2006a, Weisenberger *et al.* 2006, Ogino *et al.* 2007) This panel was shown to give an accurate and specific estimation of the genomic status of CIMP (Ogino *et al.* 2007). Fourthly, to characterise CIN using high resolution genome-wide array CGH. This systematic approach was used to provide very detailed information on the molecular subtypes, in order to investigate new associations. At the beginning of this project, a total of 67 paired “normal” and tumour DNA samples were assembled; 50 samples from the Sheffield population and 17 from the Sheffield tissue bank as described in Section 2.1.3.4. Samples were chosen based on the availability of both normal and tumour DNA (or tissue).

4.2 Results

4.2.1 DNA quality

DNA extraction was performed as described in Section 2.2.1. As described in Section 2.2.2 and Section 2.2.3, samples included in the molecular classification study were checked for several quality measures; DNA concentration/quantity, 260/280 ratio (ideal ratio: 1.8-2.0), 260/230 ratio (ideal ratio: 1.8-2.2) and degree of DNA fragmentation. The latter is shown in Figure 4.1. Samples were selected if their DNA concentration was more than 50 ng/µl and if more than 2.5 µg DNA was available with acceptable 260/280 and 260/230 ratios. The 2.5 µg DNA quantity limit was the minimum amount required to perform all the molecular tests described in the following sections. Samples with very low 260/230 were rescued through sodium acetate/ethanol precipitation as described in Section 2.2.4. Samples were only included if their DNA was of high molecular weight. A total of 67 paired DNA samples (134 normal and tumour DNA samples) described above were checked for these quality measures. Out of the 50 Sheffield population samples, 10 did not qualify for inclusion. However, all of the 17 tissue bank samples were included. A total of 57 samples were carried forward to the next step.

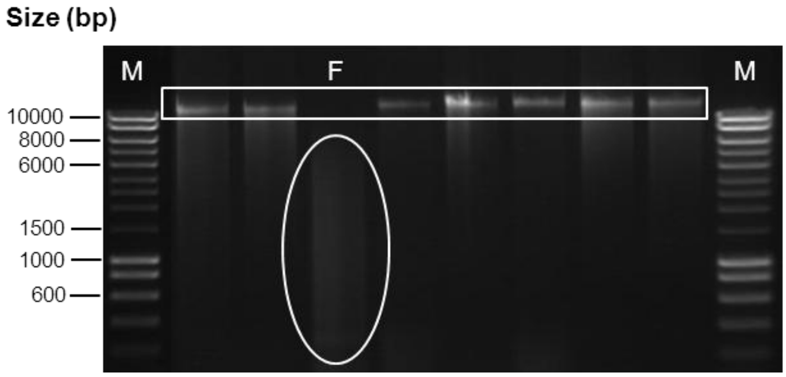


Figure 4.1 DNA fragmentation test on agarose gel

Lane F represents an example of a sample with fragmented DNA (circled) in comparison to samples with high molecular weight DNA (in rectangle). Fragmented DNA samples were excluded from the study. M: Hyperladder I.

4.2.2 Matching tumour and normal DNA test

After sample quality validation, the next step of the molecular classification was to confirm that the normal and tumour DNA samples matched (i.e. were from the same patient). This confirmation step was done by comparing two highly polymorphic pentanucleotide markers, termed penta C and penta D. These markers were included in the MSI analysis kit (Section 4.2.3), thus, the matching test was performed as part of the MSI test. Two main reasons lie behind the choice of penta C and penta D to demonstrate any sample mix-up; firstly, the combined probability that penta C and penta D (based on their allele frequency) will both randomly match between two unrelated individuals is 0.2-0.6% (depending on the population), secondly, penta-nucleotide markers are stable in mismatch repair deficient cells (Bacher *et al.* 2004). In total, 57 pairs of matching normal and tumour DNA were tested using the MSI analysis kit. Comparing penta C and penta D data, 3 pairs of samples, 2 from the Sheffield population (CA147 and CA182) and 1 from Sheffield tissue bank sample set (CA945), were not matched and thus were excluded from the study. An example of a mismatched pair of DNA samples is given in Figure 4.2.

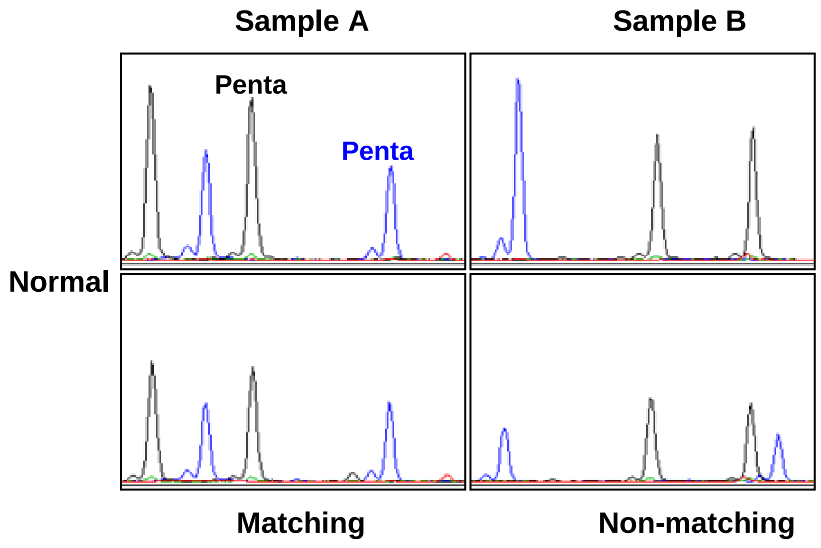


Figure 4.2 Matching normal and tumour DNA test

4.2.3 Microsatellite instability analysis

The MSI status of the matching "normal" and tumour DNA samples was determined using the commercially available QFPCR kit: MSI Analysis System Version 1.2 Kit (Section 2.2.8). The kit includes 5 quasimonomorphic mononucleotide microsatellite markers; BAT-25, BAT-26, NR-21, NR-24 and MONO-27, and the two highly polymorphic pentanucleotide markers penta C and penta D (Figure 4.3). The MSI markers were considered unstable if they exhibit a change in size in the tumour DNA compared to the paired normal DNA. The samples were classified to have high MSI (MSI-H) if they showed instability in two or more of the 5 mononucleotide markers, low MSI (MSI-L) if they showed instability in one of the markers, and MSS if none of the markers showed instability (Murphy *et al.* 2006).

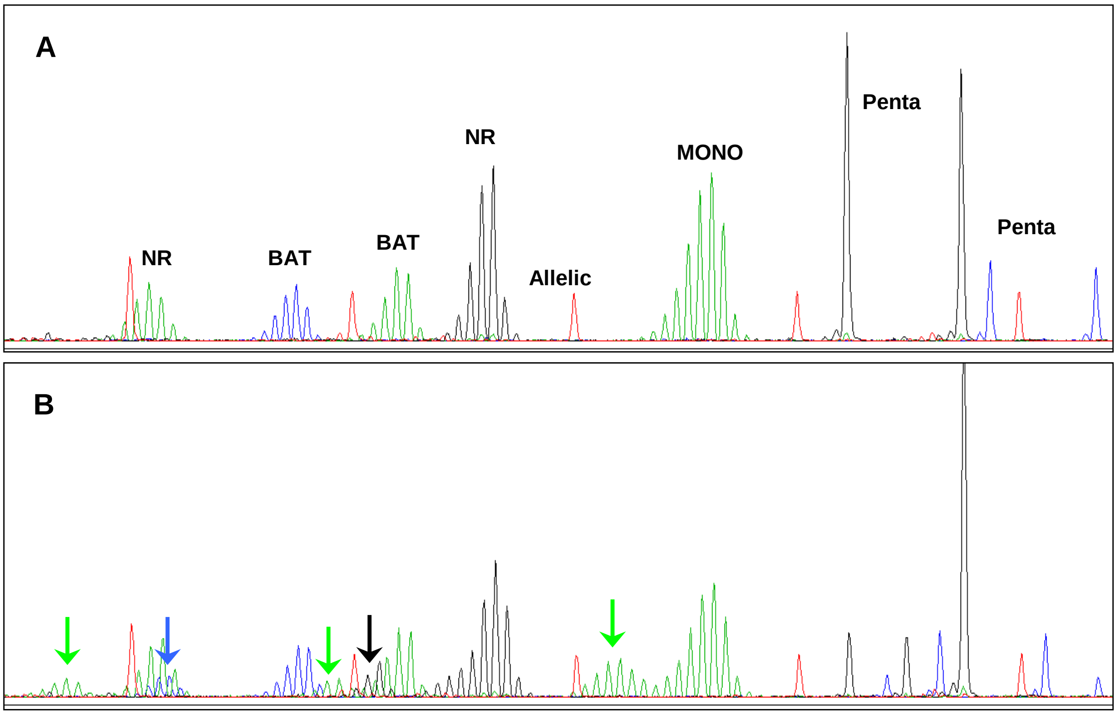


Figure 4.3 MSI analysis system v2.1

Panel A represents a normal MSS sample. Panel B represents an MSI-H sample. Marked with coloured arrows are the size aberrations of the quasimonomorphic mononucleotide markers. For a sample to be considered MSI-H, two or more markers should show size aberration. Penta C and D are pentanucleotide markers used to verify that the matching normal/tumour DNA samples were from the same patient. The MSI-H sample presented is a mix of DNA from MSI-H HCT-116 and MSS normal sample, hence, the several penta C and penta D peaks.

4.2.3.1 MSI kit sensitivity testing

In order to evaluate the sensitivity of the MSI kit, serial dilutions of MSI-H DNA were mixed with MSS normal DNA (10%, 30% and 50% MSI-H). The MSI-H DNA was extracted from HCT-116 cells which carry a homozygous *MLH1* mutation (c.755C>A) and is known to exhibit MSI-H (http://www.sanger.ac.uk/genetics/CGP/CellLines/). The mixed DNA samples were then tested using the MSI kit. As shown in Figure 4.4, microsatellite markers showing instabilities were still detected (with careful examination) at a level of 10% of the total DNA.

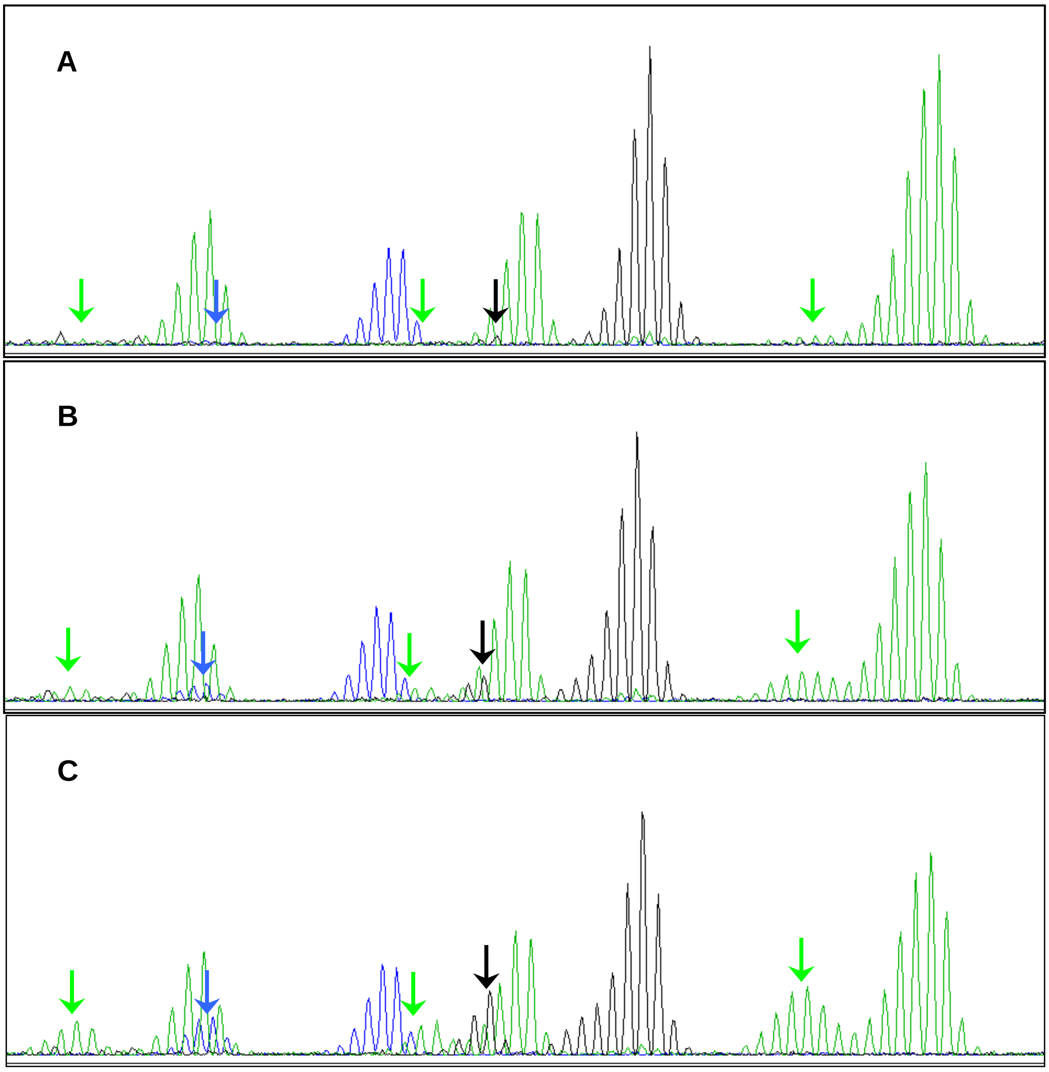


Figure 4.4 MSI kit sensitivity testing

MSI kit sensitivity testing was performed using serial dilutions of MSI-H DNA mixed with MSS DNA. A) 10% dilution, B) 30% dilution and C) 50% dilution, the MSI-H status was still detectable at the 10% dilution with careful examination

4.2.3.2 MSI results

Two of the tested CRC samples (n=57) were found to be MSI-H (Figure 4.5). The first MSI positive sample, CA945, was determined to have high levels of MSI (Figure 4.5 A), however, it is one of the samples that were considered to be contaminated in Section 4.2.2 as it had 3 alleles of both penta C and penta D. Comparing the patterns of the extra alleles with all the other DNA samples analysed within the same run excluded contamination at the MSI test level. The pattern was also compared to the samples extracted during the same run and it also excluded contamination on the DNA extraction level. Repeating the MSI test confirmed the presence of the extra alleles. The second MSI positive sample, CA008, had very low levels of MSI instability. According to the sensitivity test, this sample contained 10% or less MSI positive DNA. Therefore, it was either a heterogeneous tumour DNA sample, or the tumour sample was highly contaminated with normal DNA. A new cut from the tumour tissue has shown that the tumour is actually highly interspersed with normal cells. Therefore, CA008 and CA945 were excluded from further examination and the remaining cases were considered to be MSS.

In summary, 4 samples were excluded at the MSI testing stage (including the matching “normal” and tumour DNA test) for either being non-matching (CA147, CA182 and CA945) or being highly interspersed with normal DNA (CA008). Therefore, out of the 57 paired DNA samples with sufficient amounts, 53 MSS cases were further characterised; 38 samples from the Sheffield population and 15 from the Sheffield tissue bank cohort. The clinical and pathological characteristics of these samples are summarised in Table 4.1.

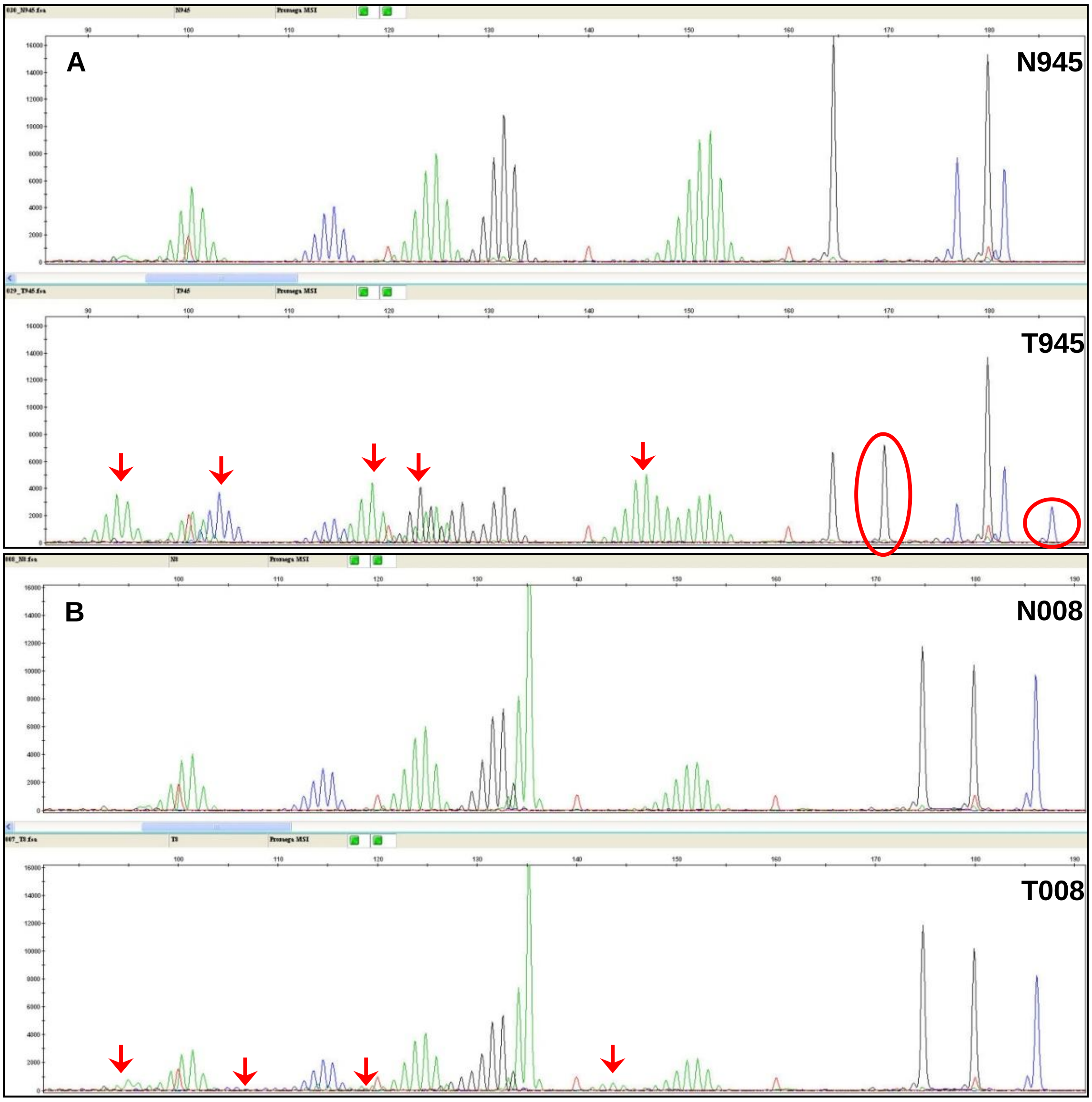


Figure 4.5 **MSI positive cases**

Represents the 2 MSI samples identified in the project. A) CA945 has high levels of MSI (marked by red arrows), however, penta C and D are indicating DNA contamination (Extra alleles marked by red circles). B) CA008 has low levels of MSI ~10% or less (marked by red arrows) high normal DNA presence.

Table 4.1 Patient details and tumour characteristics of the paired samples

		Sheffield population	Tissue Bank
Median age (range)/years		67 (51-87)	76 (41-90)
Gender (n (%))			
	Male	21 (55.3)	8 (53.0)
	Female	17 (44.7)	7 (47.0)
Tumour location (n(%))			
	Proximal	8 (21.1)	11 (73.3)
	Distal	10 (26.3)	1 (6.7)
	Rectal	20 (52.6)	3 (20.0)
Duke's Stage (n(%))			
	A	2 (5.3)	2 (13.3)
	B	11 (28.9)	3 (20.0)
	C	22 (57.9)	9 (60.0)
	D	1 (2.6)	0 (0.0)
	Unknown	2 (5.3)	1 (6.7)
Family history CRC (n(%))*			
	None	26 (68.4)	NA
	1 relative	9 (23.7)	
	≥ 2 relatives	3 (7.9)	
Metastastatic (n(%))			
	Yes	10 (26.3)	NA
	No	26 (68.4)	
	Unknown	2 (5.3)	
Differentiation (n(%))			
	Well/Moderate	11 (28.9)	1 (6.7)
	Moderate	24 (63.2)	8 (53.3)
	Poor	2 (5.3)	5 (33.3)
	Unknown	1 (2.6)	1 (6.7)
Survival			
	Alive	13 (34.2)	NA
	Dead	24 (63.2)	
	Unknown	1 (2.6)	
Median survival (range)/years		2.54 (0.08-7.58)	NA

* Family history includes first degree relatives only

4.2.4 Mutations in *APC*, *TP53*, *KRAS*, *BRAF* and *PIK3CA*

The next step of the CRC molecular classification was to determine the mutational status of *APC*, *TP53*, *KRAS*, *BRAF* and *PIK3CA* genes, which are known to play a key role in the development of sporadic CRC through the different molecular pathways (Figure 1.5). Somatic mutations were investigated in the 53 matching normal/tumour DNA samples that passed the quality and MSI testing.

4.2.4.1 *APC* gene mutation cluster region sequencing

Inactivating mutations in the TSG *APC* are one of the earliest genetic events that initiate CRC development (Kinzler and Vogelstein 1996, Kohler *et al.* 2008). Somatic

APC mutations were previously shown to occur in 34-70% of sporadic CRC cases (Miyaki *et al.* 1994, Luchtenborg *et al.* 2004). The vast majority of the *APC* somatic mutations occur in the mutation cluster region (MCR) in exon 15 (Miyoshi *et al.* 1992, Luchtenborg *et al.* 2004). The MCR and its surrounding region (codons 1265-1560) in *APC* exon 15 were sequenced in the screened cohort. Of the 53 cases, 29 (54.7%) had previously reported somatic mutations in the MCR (Table 4.2). All the mutations were only found in the tumour DNA (not present in normal DNA) and they were all confirmed by bi-directional sequencing. Moreover, all the somatic mutations were either frameshift (71%) or nonsense (29%) and they were predicted to result in truncated APC protein. This pattern of mutations agrees with previously published data (Beroud and Soussi 1996). None of the cases were found to carry more than one mutation.

APC deletion results were available through the aCGH data (Section 4.2.6). Out of the 53 cases analysed by aCGH, 9 cases were shown to have deletions in the *APC* region, 3 of these samples did not have a somatic mutation in the MCR (Table 4.2). Therefore, a total of 32 cases (60.4%) had at least 1 genetic defect in the *APC* gene.

Table 4.2 ***APC*** **somatic mutations**

Sample	Mutation	Codon	Type	*APC* deletions (aCGH)
CA828	c.3871C>T	1291	Nonsense	Del
CA206	c.3883G>T	1295	Nonsense	None
CA096	c.3970C>T	1303	Nonsense	None
CA037	c.3916G>T	1306	Nonsense	None
CA104	c.3916G>T	1306	Nonsense	None
CA218	c.3927_3931del5	1309	Frame shift	None
CA045	c.3933_3934ins1	1312	Frame shift	None
CA085	c.3964G>T	1322	Nonsense	None
CA097	c.3982C>T	1328	Nonsense	None
CA249	c.4033G>T	1345	Nonsense	Del
CA098	c.4037C>G	1346	Nonsense	None
CA079	c.4054_4062dup7	1354	Frame shift	Del
CA138	c.4127_4128del2	1376	Frame shift	None
CA150	c.4216C>T	1406	Nonsense	None
CA244	c.4271del1	1424	Frame shift	Del
CA080	c.4271_4280del10	1424	Frame shift	None
CA184	c.4303A>T	1435	Nonsense	None
CA023	c.4328_4329insC	1443	Frame shift	None
CA078	c.4342del1	1448	Frame shift	None
CA109	c.4350C>T	1450	Nonsense	None
CA153	c.4350C>T	1450	Nonsense	None
CA741	c.4350C>T	1450	Nonsense	Del
CA795	c.4391_4394delAGAG	1464	Frame shift	None
CA863	c.4391_4394delAGAG	1464	Frame shift	Del
CA221	c.4415del1	1472	Frame shift	None
CA112	c.4421del1	1474	Frame shift	None
CA088	c.4582_4603del22	1528	Frame shift	None
CA114	c.4666_4667ins1	1554	Frame shift	None
CA208	c.4666_4667ins1	1554	Frame shift	None
CA213	None	N/A	N/A	Del
CA142	None	N/A	N/A	Del
CA158	None	N/A	N/A	Del

Table 4.2 summarises all the somatic mutations and genomic deletions (Section 4.2.6) that were identified in *APC* in the screened cohort. The mutations are in order of location. Twenty one cases did not have any *APC* defects.

4.2.4.2 *TP53* gene sequencing

Mutations in *TP53* represent the most common genetic alteration in human cancers. According to the International Agency for Research on Cancer (IARC) *TP53* database, CRC has the 2nd highest level of *TP53* mutations (43.3%) (http://www-p53.iarc.fr/index.html, accessed June, 2011). Mutations in *TP53* are spread across the gene, however, most mutations occur in the region between exon 4 and exon 9

(>98% according to IARC *TP53* database). *TP53* exons 4-9 and their exon/intron boundaries were sequenced in this study. Previously reported somatic *TP53* mutations were identified in 49% (n=26) of the screened cohort (Table 4.3). One case, CA208, carried 2 somatic mutations, thus, the number of the identified mutations was 27 (Table 4.3). Most of these somatic mutations were missense (70.4%) and the rest were nonsense (14.8%), frameshift (11.1%) and splice site (3.7%). This pattern of *TP53* mutations agrees with the published data (http://www-p53.iarc.fr/index.html, accessed June, 2011). Moreover, most of these mutations occurred in exons 7 and 8 (~63%) which also agrees with the published figures.

Throughout the sequenced regions of *TP53*, 7 polymorphic SNPs were identified. These SNPs enabled basic LOH analysis by comparison of variable alleles between the tumour and the normal DNA (Figure 4.6). Out of the 53 screened cases, 27 cases (51%) were informative (i.e. had at least 2 polymorphic SNPs) and the LOH status was therefore determined. For the rest of the cases (49%, n=26), there were fewer than 2 polymorphic SNPs, therefore, they were considered uninformative and their LOH status was undetermined. Out of the 27 informative cases, 16 cases (30.2% of the total cases) showed evidence of *TP53* LOH. Interestingly, one of the CRC cases (CA045) had a germline mutation (c.935C>G) in *TP53* exon 9 (Figure 4.7). Surprisingly, this mutation was on the allele that showed LOH (Figure 4.7 and Figure 4.8). The sample also had a somatic mutation (c.524G>A) in exon 5 on the other allele (Figure 4.9). Interestingly, patient CA045 was not diagnosed with CRC at an early age (80 years); however, she had three 1st degree relatives with cancer.

TP53 deletion information was also available through the analysis of the aCGH data (Section 4.2.6). Out of the 53 cases analysed by aCGH, 17 carried *TP53* deletions. Comparing the aCGH results with the 27 LOH informative cases showed that 22 cases were concordant and 5 were not (CA79, CA86, CA249, CA271 and CA828).

These 5 discordant samples were shown to have LOH by SNP analysis, however, they were found to be diploid by aCGH analysis. This result can be explained by mono-allelic inheritance from the parental cancerous cell or gene conversion events. In summary, 31 cases (~58.5%) had at least one defect in *TP53* (Table 4.3).

Table 4.3 *TP53* somatic mutations

Sample	Mutation	Exon	Codon	Type	Amino acid change	Deletion/LOH
CA088	c.250del1	4	84	Frame shift	N/A	Del
CA206	c.455-456ins1	5	152	Frame shift	N/A	LOH
CA045	c.524G>A	5	175	Missense	p.R175H	Del/LOH
CA098	c.524G>A	5	175	Missense	p.R175H	Del
CA249	c.524G>A	5	175	Missense	p.R175H	LOH
CA828	c.524G>A	5	175	Missense	p.R175H	LOH
CA090	c.559+1G>T	i5-intron	NA	Splice site	N/A	None
CA863	c.586C>T	6	196	Nonsense	N/A	Del/LOH
CA053	c.592G>T	6	198	Nonsense	N/A	Del/LOH
CA097	c.637C>T	6	213	Nonsense	N/A	Del/LOH
CA158	c.722C>G	7	241	Missense	p.S241C	None
CA142	c.725G>T	7	242	Missense	p.C242F	None
CA208	c.730G>A	7	244	Missense	p.G244S	None
CA632	c.733G>A	7	245	Missense	p.G245S	Del/LOH
CA104	c.733G>A	7	245	Missense	p.G245S	Del/LOH
CA218	c.733G>A	7	245	Missense	p.G245S	Del/LOH
CA213	c.751A>C	7	251	Missense	p.I251L	None
CA223	c.811G>A	8	271	Missense	p.E271K	Del
CA086	c.818G>A	8	273	Missense	p.R273H	LOH
CA244	c.817C>T	8	273	Missense	p.R273C	Del/LOH
CA153	c.844C>G	8	282	Missense	p.R282G	None
CA079	c.844C>T	8	282	Missense	p.R282W	LOH
CA184	c.844C>T	8	282	Missense	p.R282W	Del
CA208	c.844C>T	8	282	Missense	p.R282W	None
CA271	c.844C>T	8	282	Missense	p.R282W	LOH
CA1120	c.847_859del13	8	283	Frame shift	N/A	None
CA085	c.916C>T	8	306	Nonsense	N/A	None
CA114	None	N/A	N/A	N/A	N/A	Del/LOH
CA201	None	N/A	N/A	N/A	N/A	Del
CA203	None	N/A	N/A	N/A	N/A	Del
CA214	None	N/A	N/A	N/A	N/A	Del
CA221	None	N/A	N/A	N/A	N/A	Del/LOH

Table 4.3 summarises all the somatic mutations, LOH and genomic deletions (Section 4.2.6) that were identified in *TP53* in the screened cohort. The mutations are in order of location. Thirty one had at least 1 defect in *TP53* (1 case with 2 somatic mutations, 18 cases with a somatic mutation and LOH/deletion, 5 cases with LOH/deletion only and 7 cases with a somatic mutation without LOH/deletion or LOH status undetermined).

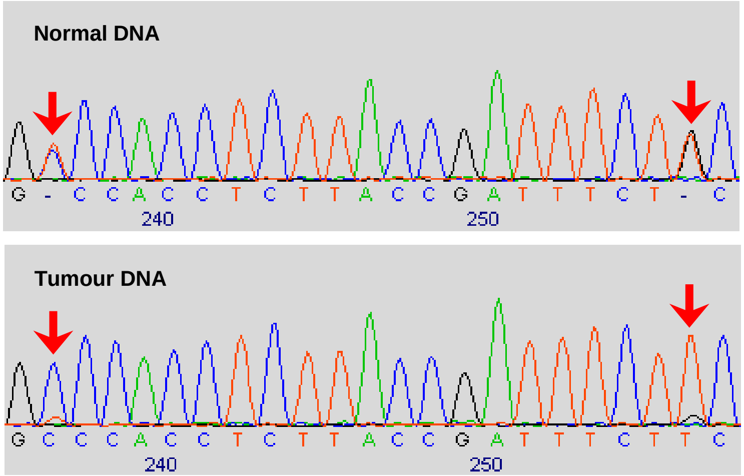


Figure 4.6 **LOH analysis**

Samples were analysed for using polymorphic SNPs comparison between normal and tumour DNA. The SNPs are marked with red arrows. As seen in the figure, the heterozygous SNPs in the normal DNA appear homozygous in the tumour DNA due to loss of one of the alleles. Traces of the normal DNA can be seen in the tumour DNA.

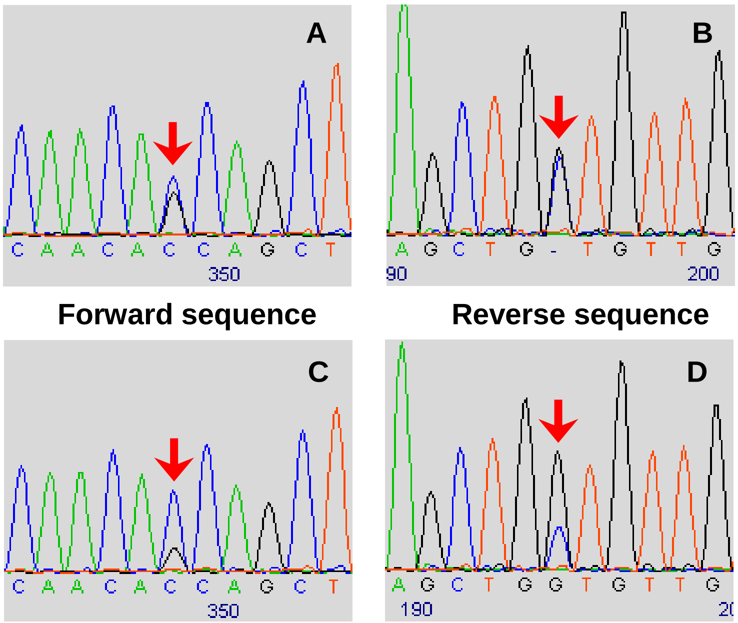


Figure 4.7 ***TP53* germline mutation c.935C>G in CA045**

Bidirectional sequencing of CA045 *TP53* exon 9 region showing the germline mutation c.935C>G in the normal DNA (A and B) and tumour DNA (C and D). As seen in the figure, the tumour DNA was considered wild-type at this position, however, traces of the germline mutation can be seen in the tumour DNA.

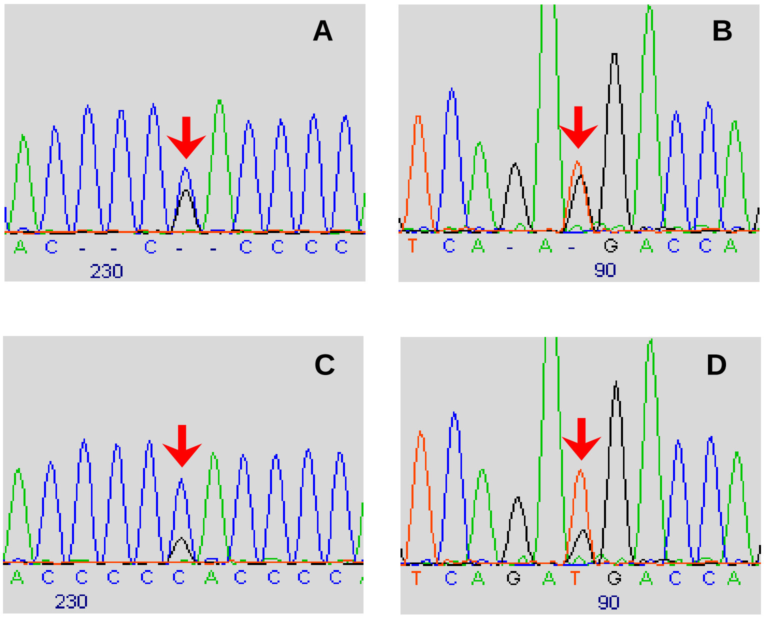
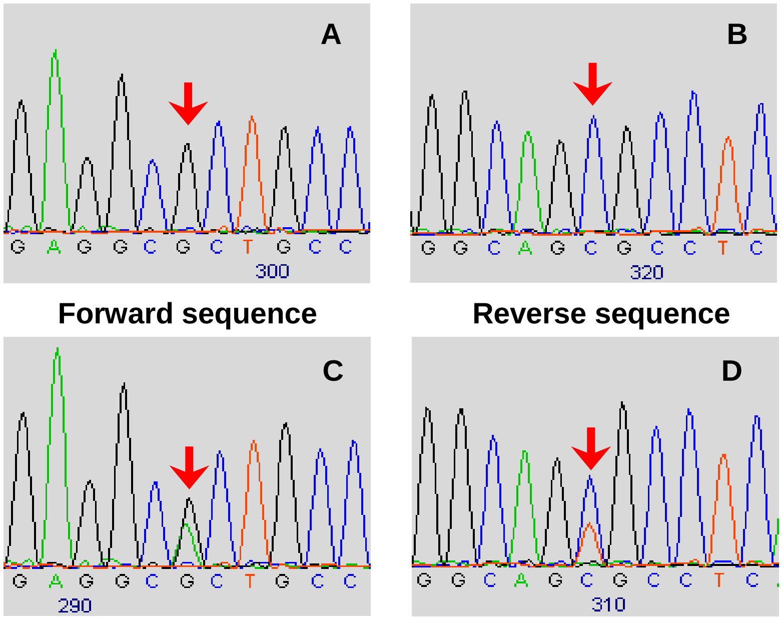


Figure 4.8 ***TP53*** **LOH in CA045**

Bidirectional sequencing of CA045 *TP53* showing 2 different heterozygous SNPs in the normal DNA (A and B) and the tumour DNA (C and D). As seen in the figures, the 2 SNPs indicate LOH in the tumour DNA.



Forward sequence

Reverse sequence

Figure 4.9 ***TP53*** **somatic mutation c.524G>A in CA045**

Bidirectional sequencing of CA045 *TP53* exon 5 region showing the somatic mutation c.524G>A in the tumour DNA (C and D). The normal DNA did not show any evidence of mutation at this position (A and B).

4.2.4.3 *KRAS* exon 1 sequencing

KRAS oncogene activation by somatic mutations has been shown to occur in 30-60% of CRC cases (Brink *et al.* 2003). Approximately, 95% of these mutations occur in exon 1 (codons 12 & 13) and ~5% in exon 2 (codon 61). Therefore, *KRAS* exon 1 was sequenced in our 53 CRC cases. Previously published somatic mutations were found in 36% (n=19) of the sequenced samples (Table 4.4). As expected, most of the mutations (84%, n=16) were in codon 12. All the mutations were only found in the tumour DNA (not present in normal DNA) and they were all confirmed by bi-directional sequencing.

Table 4.4 *KRAS* somatic mutations

Sample	Mutation	Codon	Type	Amino acid change
CA037	c.35G>T	12	Missense	p.G12V
CA046	c.35G>A	12	Missense	p.G12D
CA080	c.35G>A	12	Missense	p.G12D
CA086	c.35G>A	12	Missense	p.G12D
CA088	c.35G>T	12	Missense	p.G12V
CA097	c.34G>T	12	Missense	p.G12V
CA104	c.35G>T	12	Missense	p.G12V
CA107	c.35G>A	12	Missense	p.G12D
CA109	c.35G>T	12	Missense	p.G12V
CA114	c.38G>A	13	Missense	p.G13D
CA135	c.38G>A	13	Missense	p.G13D
CA138	c.35G>C	12	Missense	p.G12A
CA153	c.34G>T	12	Missense	p.G12V
CA201	c.35G>T	12	Missense	p.G12V
CA208	c.35G>T	12	Missense	p.G12V
CA249	c.35G>T	12	Missense	p.G12V
CA632	c.38G>A	13	Missense	p.G13D
CA824	c.35G>C	12	Missense	p.G12A
CA1350	c.34G>T	12	Missense	p.G12V

Table 4.4 summarises all the somatic mutations that were identified in *KRAS* in the screened cohort. The samples are in numerical order.

4.2.4.4 *BRAF* exon 15 sequencing

Activating mutations in the *BRAF* oncogene were previously shown to be involved in the development of a wide range of human cancers including CRC (Davies *et al.* 2002). However, *BRAF* somatic mutations are relatively rare (~12%) in CRC. Nonetheless, they usually tend to cluster within the MSI-H pathway where they were observed with a frequency of 40-50% compared to ~5% only in MSS CRC (Oliveira *et al.* 2007, Shen *et al.* 2007). Most of the *BRAF* mutations occur in exons 11 and 15. Approximately 80% of these mutations occur in the mutational hotspot in exon 15 specifying amino acid V600. Therefore, exon 15 was sequenced and somatic mutations in *BRAF* were identified in 2 cases (3.8%) (Table 4.5). One of the mutations was in the hotspot codon encoding V600E. The second mutation is a relatively rare, but known, mutation (N581S). All the mutations were only found in tumour DNA (not present in normal DNA) and they were confirmed by bi-directional sequencing.

Table 4.5 *BRAF* somatic mutations

Sample	Mutation	Codon	Type	Amino acid change
CA208	c.1742A>G	581	Missense	p.N581S
CA271	c.1799T>A	600	Missense	p.V600E

Table 4.5 summarises all the somatic mutations that were identified in *BRAF* in the screened cohort. The samples are in numerical order.

4.2.4.5 *PIK3CA* exons 9 and 20 sequencing

Activating mutations in the *PIK3CA* oncogene were previously shown to play an important role in the development of several human cancers including CRC (Samuels *et al.* 2004). Somatic *PIK3CA* mutations were reported in ~13-30% of sporadic CRC cases with mutational hotspots in exons 9 and 20 (~80% of the mutations) (Samuels *et al.* 2004, Ikenoue *et al.* 2005, Velho *et al.* 2005). Therefore, exons 9 and 20 were sequenced in this study and previously identified somatic mutations were found in 6 cases (11.3%) (Table 4.6), consistently, most of which (n=5) were in exon 9, as previously reported (Samuels *et al.* 2004). All the mutations were only found in tumour DNA (not present in normal DNA) and they were confirmed by bi-directional sequencing.

Table 4.6 *PIK3CA* somatic mutations

Sample	*PIK3CA*	Exon	Codon	Type	Amino acid change
CA023	c.1633G>A	9	545	Missense	p.E545K
CA046	c.1633G>A	9	545	Missense	p.E545K
CA107	c.1633G>A	9	545	Missense	p.E545K
CA153	c.1633G>A	9	545	Missense	p.E545K
CA795	c.1633G>A	9	545	Missense	p.E545K
CA109	c.3140A>G	20	1047	Missense	p.H1047R

Table 4.6 summarises all the somatic mutations that were identified in *PIK3CA* in the screened cohort. The mutations are in order of genomic location.

4.2.5 CpG island methylator phenotype

As described in section 1.3.3.3, CpG island methylator phenotype (CIMP) occurs in 20-40% of CRC cases and plays an essential role in sporadic CRC development (Weisenberger *et al.* 2006, Ogino *et al.* 2007, Goel and Shin 2008). The MS-MLPA kit ME042-A1 (MRC-Holland) was used to determine the methylation status of the 53 normal/tumour paired DNA samples (Section 2.2.12). Eight methylation markers were included in the kit (*CACNA1G*, *CDKN2A*, *CRABP1*, *IGF2*, *MLH1*, *NEUROG1*, *RUNX3* and *SOCS1*), however, *SOCS1* was excluded from the analysis because only one probe was available. As described in section 2.2.12.4.3, no CIMP (CIMP-N) was defined as an absence of any methylated markers, CIMP low (CIMP-L) as the presence of 1-4 methylated markers and CIMP high (CIMP-H) as 5 or more (Ogino *et al.* 2007, Barault *et al.* 2008). During our study, the ME042-A1 MS-MLPA kit was still under development by MRC-Holland and not commercially available. Therefore, it was decided to attempt to validate the kit in-house.

4.2.5.1 Pyrosequencing, MS-MLPA and testing methylation controls

In order to validate the MS-MLPA kit, a commercially available universally methylated DNA (UMD) (CpGenome) and HCT116 cell line DNA were used as positive controls. In theory, UMD should give ~100% methylation in all of the markers; on the other hand, HCT116 cells which are MSI-H and CIMP-H, have positive and negative markers corresponding to various levels of methylation (Hinoue *et al.* 2009). One normal DNA sample from case CA037 was also used as negative control. As the gold standard technique for DNA methylation studies, pyrosequencing was used to test the positive and negative controls for three of the markers (*MLH1*, *NEUROG1* and *CRABP1*), which are available in the MS-MLPA kit. *MLH1* primers were purchased from Qiagen, while *NEUROG1* and *CRABP1* primers were designed as described in Section 2.2.14.1, and obtained from Applied Bio-systems. The primers were designed to target areas previously investigated in CIMP studies (Ogino *et al.* 2007). The PyroMark assay design software (v2.0) assigns quality measures for the designed primers and assays as high, medium or low. Both *CRABP1* and *NEUROG1* designs scored “high” on the quality measures.

Following bisulphite treatment, UMD, HCT116 and CA037 DNA samples were analysed by pyrosequencing (“pyrosequenced”) to confirm their methylation status. As described in Section 2.2.13, DNA bisulphite treatment results in the conversion of unmethylated cytosines into uracil with no effect on methylated cytosines. Therefore, all cytosines that are not followed by guanines should, in theory, be fully converted into uracil following bisulphite treatment, regardless of the methylation status of the sample. Uracil nucleotides will subsequently be replaced by thymidines during pyrosequencing. Therefore, a cytosine that is not followed by a guanine is usually selected as a marker to measure the bisulphite treatment quality (Figure 4.10). High quality bisulphite treatment should result in a thymdine instead of cytosine peak at the position indicated. Likewise, unmethylated cytosines followed by guanines (CpG) will also be converted into uracil and thus appear as thymidines in the pyrogram (Figure 4.11 A). On the other hand, methylated cytosines followed by guanines (CpG) will be protected and will remain as cytosines in the final pyrogram (Figure 4.11 B). A ratio of the thymidines to cytosines at each CpG position will reflect the methylation level of the DNA. Usually, several CpGs are used to test the methylation of a specific marker or gene. An average of these CpGs represents the methylation level of the marker.

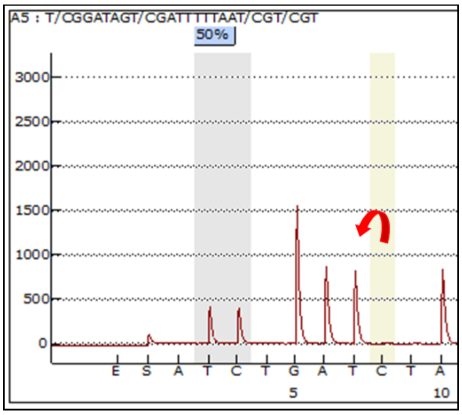


Figure 4.10 Pyrogram showing a bisulphite treatment control

A pyrogram representing an unmethylated cytosine (highlighted in a light yellow box) followed by adenine. The cytosine was fully converted into thymidine as indicated by the red arrow. Highlighted in light blue box, a 50% methylated CpG.

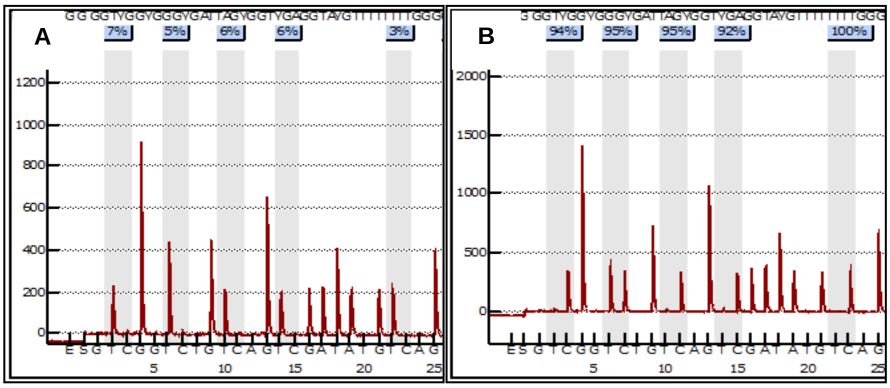


Figure 4.11 Unmethylated and methylated pyrograms

Shaded in light blue, cytosine nucleotides of the target CpGs. The small boxes on top of each CpG represent the calculated methylation percentage at each location. Blue coloured boxes indicate high pyrosequencing quality. A) Unmethylated sample, B) Highly methylated sample.

The 3 methylation controls (UMD, HCT116 and CA037) were bisulphite treated and then pyrosequenced to confirm their methylation status. The methylation levels of the 3 controls were as expected (Table 4.7). The controls were then tested using the MS-MLPA kit. As described in Section 2.2.12 and, as summarised in Figure 2.5, the methylation sensitive enzyme *HhaI* is used in the MS-MLPA digestion/ligation step. In the case of a methylated DNA sample, the MS-MLPA probes will ligate to methylated DNA sequences and will be protected against digestion. Thus, these regions will be amplified in the PCR step, and produce a signal during capillary electrophoresis appearing as peaks when analysed by the genemapper software (Figure 4.12 B).

In the case of an unmethylated DNA sample, MS-MLPA probes will ligate to unmethylated DNA sequences, and hence, they will be digested by *HhaI*. Consequently, they will fail to amplify and no peaks will be identified (Figure 4.12 C). The ratio between methylation sensitive peak areas and reference peak areas represents the methylation level. This is calculated using a Microsoft Excel based spreadsheet and in-house algorithm described in Section 2.2.12.4. The final quantitative estimates of the methylation levels are presented as barcharts (Figure 4.13). Table 4.7 summarises both the pyrosequencing and the MS-MLPA results for UMD, HCT116 and CA037. As demonstrated in the table, there is very good concordance between pyrosequencing and MS-MLPA in determining the methylation levels of these controls. The results of the controls were confirmed twice by pyrosequencing.

Table 4.7 Positive & negative controls using pyrosequencing & MS-MLPA

	Methylation levels (%)					
	Pyrosequencing			**MS-MLPA**		
Sample	***MLH1***	***NEUROG1***	***CRABP1***	***MLH1***	***NEUROG1***	***CRABP1***
CA037	14.5	7.0	5.4	0.5	9.3	4.4
HCT116	11.5	96.3	94.6	0.4	100.0	95.9
UMD	96.3	89.5	94.4	94.9	97.2	95.1
No of CpGs*	4	8	7	3	6	4

*Number of CpGs analysed for each marker, the methylation levels of the markers are determined by averaging the CpGs for each marker.

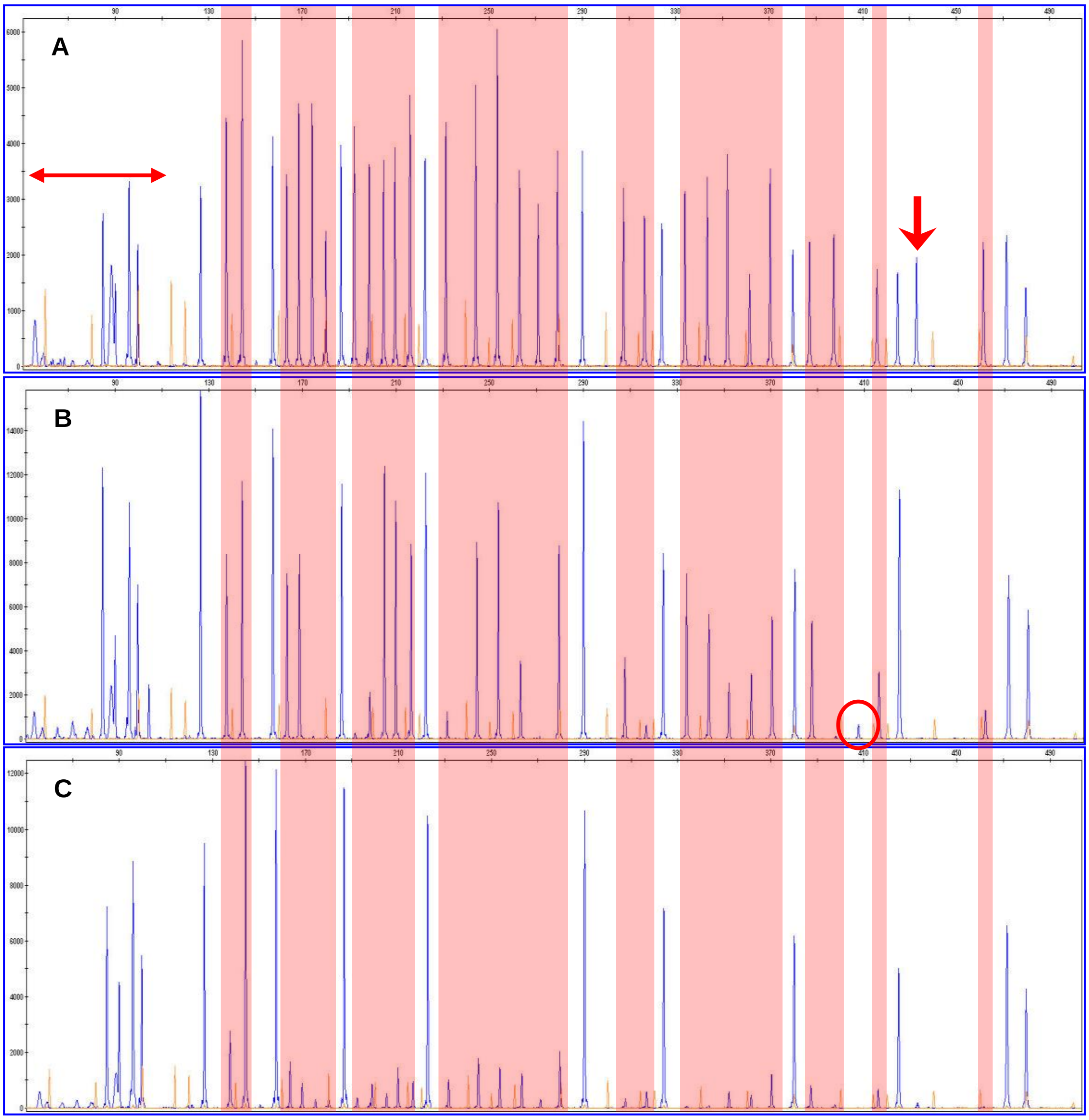


Figure 4.12 MS-MLPA genemapper peaks

Peaks highlighted in light red boxes are the products of methylation sensitive probes. Peaks in between them represent the products of the reference probes. Peaks marked by the horizontal red arrow are the products of the QC metrics and sex determining probes. Peak marked by the vertical red arrow is the digestion control probe. This peak should disappear in digested samples. Peak marked by the red circle is the *BRAF* V600E mutation specific peak. This peak will appear only in the case of the mutation. A) Non-digested N212 DNA sample B) Methylated T271 DNA sample with *BRAF* V600E mutation C) Unmethylated N212 sample

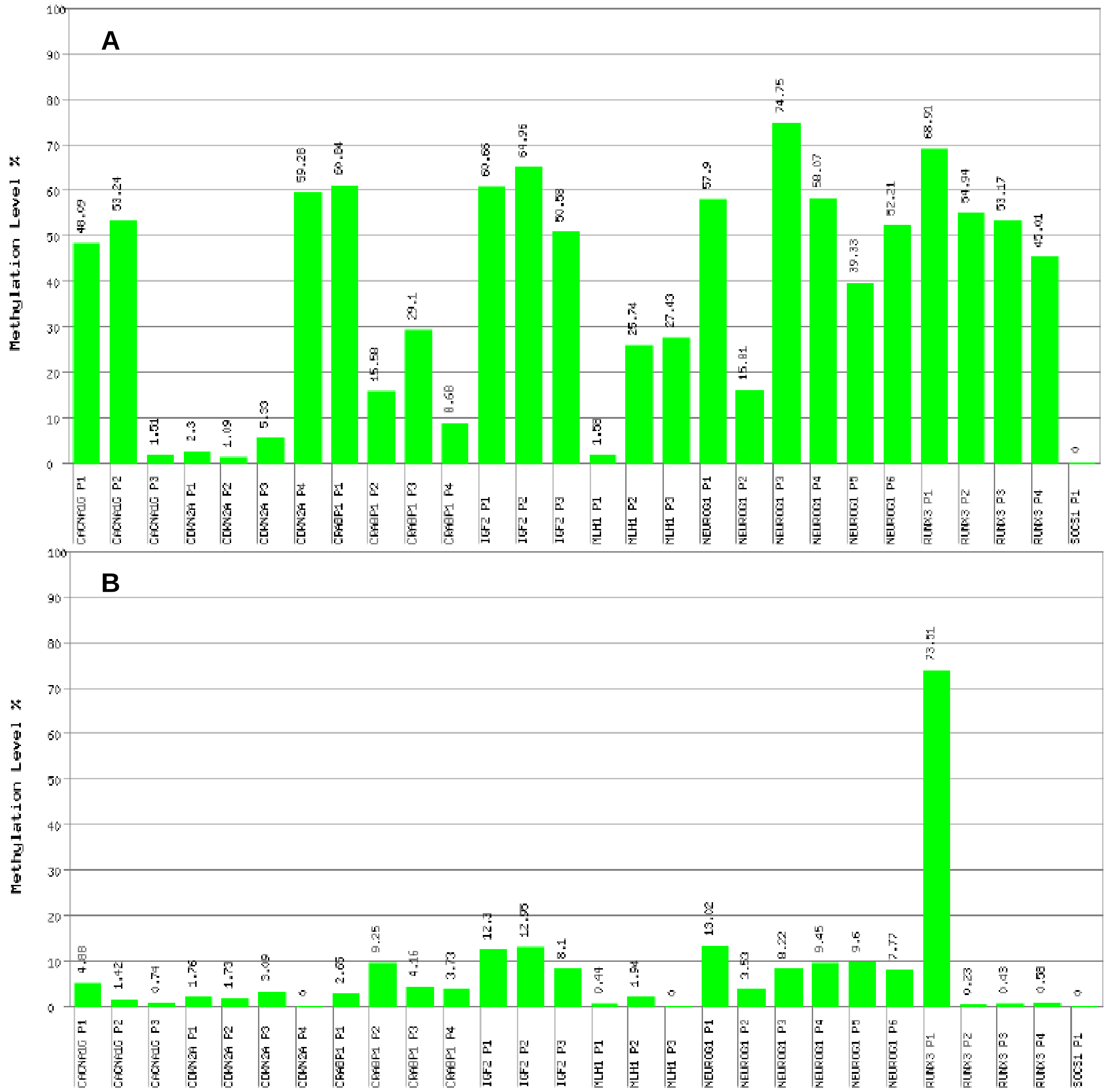


Figure 4.13 Barcharts showing MS-MLPA calculated methylation levels

A) Methylated sample T271, B) Unmethylated sample N271

4.2.5.2 MS-MLPA quantification test

After confirming the specificity of the MS-MLPA kit, its ability to produce quantifiable data was tested. MS-MLPA as technology was previously described to be semi-quantitative (Jeuken et al. 2007). In order to test the quantification of the MS-MLPA ME042-A1 kit, serial dilutions of the UMD were used. CIMP negative DNA from the tested sample CA037 was used to do serially dilute the UMD (75%, 50% and 25% UMD). The results of the serial dilution are shown in Table 4.8 and they indicate the ability of the MS-MLPA kit to quantify the methylation levels of the indicated DNA samples.

Table 4.8 MS-MLPA ME042-A1 quantification test

UMD serial dilution	**CIMP markers and average methylation levels (%)**						
	IGF2	***RUNX3***	***NEUROG1***	***MLH1***	***CDKN2A***	***CRABP1***	***CACNA1G***
100%	101.02	99.70	97.21	94.90	93.87	95.13	107.17
75%	75.55	79.67	80.77	72.74	69.54	76.37	65.68
50%	50.24	54.64	54.93	48.04	48.00	53.88	51.94
25%	35.91	31.88	38.62	30.51	32.17	35.50	35.40

4.2.5.3 MS-MLPA reproducibility

Finally, the reproducibly of the MS-MLPA kit was tested by performing duplicate experiments. As its methylation status is known, HCT116 DNA was duplicated within both the same run and in separate runs. However, patient DNA samples had their methylation levels tested by MS-MLPA and these were then repeated in different runs, to confirm the findings and to further test the reproducibility of the MS-MLPA kit. In total, HCT116 DNA was tested 5 times and 4 case DNA samples (2 CIMP-H, 1 CIMP-L and 1 CIMP-N) were analysed twice. The results of these duplications are summarised in Table 4.9.

Table 4.9 MS-MLPA reproducibility tests

	Methylation Levels (%)						
	CACNA1G	***CDKN2A***	***CRABP1***	***IGF2***	***MLH1***	***NEUROG1***	***RUNX3***
T080-1	1.6	1.3	3.4	29.9	1.0	5.7	1.7
T080-2	2.1	1.6	3.9	25.7	1.0	5.2	1.8
T184-1	7.6	1.8	10.9	50.0	0.4	46.1	2.5
T184-2	9.3	1.9	11.0	51.0	1.3	45.7	2.4
T114-1	2.1	24.7	25.5	60.3	0.0	52.2	23.7
T114-2	1.9	31.1	31.1	68.9	1.8	60.0	24.9
T271-1	34.3	17.0	28.6	58.7	18.3	49.7	51.0
T271-2	32.0	19.3	28.4	59.4	21.9	50.4	62.2
HCT116-1	95.7	38.3	95.9	100.0	0.4	100.0	4.0
HCT116-2a	88.6	41.6	98.1	100.0	0.7	100.0	6.1
HCT116-2b	89.8	35.8	99.0	100.0	2.1	100.0	5.9
HCT116-3a	77.5	37.2	92.0	100.0	2.1	97.1	7.7
HCT116-3b	85.2	41.1	100.0	99.2	1.1	100.0	5.6

Table 4.9 summarises the results of the repeated samples. All the case samples were repeated in different runs. HCT116 was either duplicated within the same run (a and b) or repeated in different runs as well.

4.2.5.4 CpG island methylator phenotype results

After validating the performance of the MS-MLPA kit, it was used to determine the methylation levels in the paired tumour and normal DNA cohort. Two main issues were identified in the methylation testing of the MS-MLPA kit and were reported back to MRC-Holland. Firstly, there was only one probe for *SOCS1*, therefore it was not included in the analysis. Secondly, *RUNX3* probe 1 had high methylation levels in the normal DNA, thus it was not included in the analysis of *RUNX3* methylation. As a final validation step, 3 more cases were confirmed using pyrosequencing. Paired DNA samples from 2 CIMP-H cases (CA114 and CA223) and a CIMP-L case (CA090) were used. Table 4.10 summarises the results of these 3 paired samples using both pyrosequencing and MS-MLPA.

As described in Section 2.2.12.4.4, sex determination probes and a specific *BRAF* V600E mutation probe (associated with CIMP-H) are part of the MS-MLPA ME042-

A1 kit. It is noteworthy here that there was a full concordance between the sex prediction of in the MS-MLPA assay and the recorded sex of each case. Moreover, the mutation status of the only *BRAF* V600E positive sample CA271 (Table 4.5) was also confirmed by the MS-MLPA assay (Figure 4.12 B).

It is noteworthy that all the MS-MLPA tests included in CIMP analysis had excellent QC metrics results (MS-MLPA QC-results were described in Section 2.2.12.4.1). The QC-tests results for all the samples are summarised in Appendix 5. Finally, out of the 53 tested samples, 6 (11.3%) were defined as CIMP-H, 41 (77.4%) as CIMP-L and 6 were defined as CIMP-N (11.3%) (Table 4.11).

Table 4.10 Random samples analysed using pyrosequencing and MS-MLPA

	Methylation levels (%)					
	Pyrosequencing			**MS-MLPA**		
Sample	***MLH1***	***NEUROG1***	***CRABP1***	***MLH1***	***NEUROG1***	***CRABP1***
N090	11.8	9.5	5.7	1.2	7.6	4.4
T090	12.5	31.0	38.3	1.2	19.8	20.4
N114	13.8	6.9	6.7	1.2	11.7	3.7
T114	13.3	45.6	54.0	0.0	52.2	23.7
N223	15.3	7.5	5.4	0.6	10.0	3.8
T223	10.8	52.5	57.0	0.7	48.3	50.8
No of CpGs	4	8	7	3	6	4

Table 4.10 summarises the reproducibility test performed on 3 paired cases. Normal DNA for the 3 samples was used to represent CIMP-N. T090 was used to represent CIMP-L and T114 and T223 were used to represent CIMP-H.

Table 4.11 Summary of the final CIMP status for all of the 53 samples

Sample	*CACNA1G*	*CDKN2A*	*CRABP1*	*IGF2*	*MLH1*	*NEUROG1*	*RUNX3*	Result
CA271	31.93	15.36	23.60	47.61	17.46	41.08	50.63	CIMP-H
CA088	31.73	45.90	38.66	56.54	0.44	56.66	39.00	CIMP-H
CA223	36.41	13.84	46.95	66.65	0.09	38.32	57.81	CIMP-H
CA109	0.05	11.48	33.55	51.76	0.57	44.09	19.56	CIMP-H
CA114	0.16	23.50	21.79	50.35	-1.21	40.51	22.64	CIMP-H
CA138	-0.02	19.97	33.84	40.84	-0.47	33.77	32.93	CIMP-H
CA795	0.36	12.56	6.86	20.92	0.13	25.86	23.12	CIMP-L
CA085	3.63	0.77	18.16	33.39	0.24	42.82	1.34	CIMP-L
CA086	0.99	0.77	17.61	57.23	10.30	40.03	13.33	CIMP-L
CA090	-0.22	34.34	15.99	52.70	0.02	12.17	2.60	CIMP-L
CA046	0.70	14.22	0.96	15.01	-0.23	12.62	1.88	CIMP-L
CA150	12.90	2.12	6.60	23.68	6.66	12.82	28.14	CIMP-L
CA244	-0.55	7.15	8.44	37.50	0.16	13.99	13.70	CIMP-L
CA249	0.35	10.09	12.82	45.25	-0.20	31.05	0.80	CIMP-L
CA023	1.06	5.89	6.59	16.07	0.94	12.30	0.08	CIMP-L
CA045	0.13	8.39	0.42	39.23	0.00	18.22	2.10	CIMP-L
CA053	1.91	0.44	-1.79	39.63	0.68	27.15	2.71	CIMP-L
CA083	-0.31	0.78	-0.55	45.26	9.87	36.73	0.65	CIMP-L
CA097	-0.74	-0.15	1.41	33.28	9.10	25.86	8.47	CIMP-L
CA104	0.98	-0.16	1.20	43.49	0.12	27.88	0.48	CIMP-L
CA107	-0.66	5.65	-2.94	40.73	0.19	32.66	-0.63	CIMP-L
CA135	-0.12	7.84	-0.07	35.97	0.57	17.68	2.15	CIMP-L
CA153	-1.10	13.72	3.27	45.75	-1.07	37.25	5.84	CIMP-L
CA167	3.00	0.86	2.36	22.87	0.53	4.92	2.64	CIMP-L
CA184	4.64	-0.28	7.34	43.57	-0.53	36.38	1.46	CIMP-L
CA201	-1.26	-1.13	-1.44	31.91	-0.37	18.74	0.93	CIMP-L
CA202	-0.33	1.91	-0.42	19.56	-0.92	13.69	0.48	CIMP-L
CA203	-0.33	0.14	1.54	37.34	0.28	15.23	-0.47	CIMP-L
CA206	-1.11	3.48	-0.35	23.23	16.00	37.45	1.59	CIMP-L
CA208	0.50	-0.52	-1.68	46.73	-1.43	19.90	1.33	CIMP-L
CA213	0.51	-0.28	0.79	41.90	14.11	9.46	2.68	CIMP-L
CA214	0.62	2.61	-1.35	15.93	0.82	17.51	0.20	CIMP-L
CA037	-0.49	1.26	-0.48	23.03	1.03	-0.42	1.13	CIMP-L
CA218	-0.54	0.01	-2.23	41.98	-0.08	1.01	3.19	CIMP-L
CA632	0.11	-0.89	-2.75	12.76	-0.35	75.08	-0.04	CIMP-L
CA741	-0.63	-0.56	12.99	9.43	0.17	-3.78	-0.57	CIMP-L
CA824	4.36	0.70	5.87	12.20	1.11	3.92	-2.07	CIMP-L
CA828	2.90	9.80	-0.56	31.80	0.18	25.29	-2.09	CIMP-L
CA863	-0.40	-0.51	-3.12	29.40	-0.48	25.09	-2.76	CIMP-L
CA110	0.56	1.32	1.63	15.04	1.24	3.25	2.79	CIMP-L
CA112	-0.30	-0.82	5.17	14.14	-0.19	-1.80	1.30	CIMP-L
CA1120	-0.02	0.11	-2.29	3.21	0.36	33.59	0.18	CIMP-L
CA1350	-0.09	0.26	8.58	55.24	-0.54	1.69	-1.67	CIMP-L
CA016	1.26	-0.36	6.68	5.59	8.67	11.99	3.80	CIMP-L
CA080	0.32	-0.71	-0.62	22.06	-0.24	-2.08	1.04	CIMP-L
CA098	-1.95	0.29	-2.37	51.57	-0.03	15.50	1.14	CIMP-L
CA142	-1.52	6.39	24.01	18.09	1.00	-7.36	1.53	CIMP-L
CA79	-0.54	-0.34	0.24	16.69	-0.32	2.56	-1.82	CIMP-N
CA122	0.63	-7.70	-0.48	-6.04	0.46	4.44	-0.78	CIMP-N
CA158	0.29	-1.69	8.22	0.10	0.24	-2.70	-2.19	CIMP-N
CA212	0.87	0.28	-0.59	10.19	0.39	-1.43	2.75	CIMP-N
CA221	-1.28	8.41	-3.96	4.77	0.27	4.55	-0.09	CIMP-N
CA248	0.71	1.42	0.52	9.72	0.13	3.26	2.02	CIMP-N

Table 4.11 summarises the methylation status of all the markers and the final methylation status for all of the samples. Red boxes represent methylated markers, green boxes represent unmethylated markers. A Result red box represents CIMP-H, orange box represent CIMP-L and green box represents CIMP-N samples respectively. The values in the boxes represent the difference in percentage between the tumour and normal DNA methylation levels. Samples are arranged according to their final methylation status.

4.2.6 Chromosomal instability

As described in section 1.3.3.1, chromosomal instability (CIN) is the most common form of genomic instability in CRC, and it accounts for between 65-85% of the cases (Pino and Chung 2010, Migliore *et al.* 2011). CIN is mainly characterised by numerical chromosomal abnormalities (aneuploidy) and structural amplifications and deletions (Pino and Chung 2010). Genome wide aCGH was the method used in this project to investigate CIN. Array CGH steps and protocol were performed according to the manufacturer's recommendations as described in Section 2.2.15. In this study, CIN was defined as a continuous variable of aberrations and samples were considered chromosomally unstable by the presence of significant amplifications or deletions in one or more region of any chromosome (Cheng *et al.* 2008).

4.2.6.1 Array comparative genome hybridisation platforms

A wide range of Agilent aCGH formats with differing resolutions are commercially available. After considering the cost of the arrays, their throughput, DNA requirement and resolution, two array formats were plausible; the whole genome 4x44K and 4x180K. The 4x refers to the number of arrays per slide and the 44K and 180K refers to the number of distinct biological features covered by the array probes (42494 and 170334 features). In terms of sample throughput and amount of DNA required, both formats were identical. In terms of resolution, the 180K platform provides an average resolution of ~13Kb based on the overall median probe spacing (11Kb in the Refseq genes). On the other hand, the 44K platform has an overall median probe spacing of 43Kb (24Kb in Refseq genes). The advantages of using the 44K platform are the lower cost and potentially lower background signals. On the other hand, the advantages of using the 180K platform are the higher resolution, which might help in detecting smaller aberrations. Moreover, with higher probe density, the 180K platform would be expected to provide higher confidence in detecting chromosomal abnormalities in mosaic and heterogeneous tumour DNA samples.

The minimum DNA input required for both arrays is 0.5 µg. Unfortunately, this amount of DNA is not available for all the tumour DNA samples included in the study. Therefore, whole genome amplification (WGA) might be necessary for some of the samples. Reviewing the literature, several publications (including Agilent's technical note) described the use of WGA DNA on aCGH with consistent and reproducible results (Brueck *et al.* 2007, Pugh *et al.* 2008). Nevertheless, other publications have shown that the use of WGA DNA on aCGH will result in the introduction of spurious aberrations (Talseth-Palmer *et al.* 2008).

4.2.6.2 Array CGH Quality control metrics results

As described in Section 2.2.15.10, the QC metrics of the CGH arrays are closely monitored to ensure high quality results. The genomic workbench software used in the analysis of the aCGH data combine the 11 QC metrics together and assigns a final GC status of either pass, marginal, fail or not applicable. All the CGH arrays included in the analysis of this study had a final QC status of pass.

4.2.6.3 Array comparative hybridisation and whole genome amplification

In order to decide upon the best platform to use (44K or 180K), and determine whether WGA affects the integrity of the results, a comparison aCGH test was performed. The test included 2 DNA samples, with and without WGA, on both array formats (Figure 4.14). Some of the QC metrics of the arrays were in the evaluate range, however, the total QC metrics score was pass and the experiment was considered valid for comparative purposes.

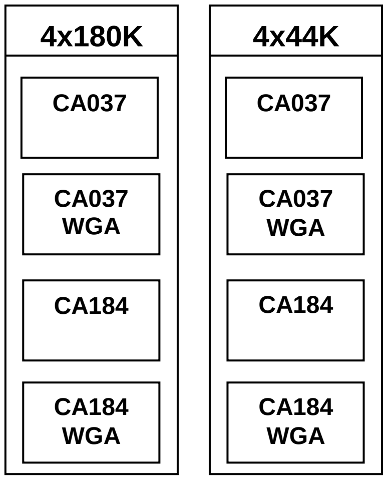


Figure 4.14 aCGH and WGA

A schematic representation of the aCGH/WGA test, CA037 and CA184 were run twice on each array, both without and without WGA.

To summarise the findings, one of the test samples was classified as chromosomally stable (CA037) and the other as unstable (CA184). CA037 results were identical with and without WGA. For the unstable CA184 sample, the direct unamplified DNA experiment was analysed first. The results of the unstable sample on both the 44K and the 180K platforms were highly comparable (Figure 4.15). In total, 7 aberrations were called on the 44K platform and 12 aberrations on the 180K platform. The aberrations on the 44K platform were all detected on the 180K platform. The missed aberrations by the 44K platform were on chrY (p and q arms), chr8p arm, chr18p arm and a small deletion on chr13. Figure 4.15 represents one of the relatively large aberrations (chr8p) missed on the 44K platform. As illustrated in the Figure, the chr8p probes on the 44K platform showed a clear shift towards the negative $\log_2$ratio, however, they did not reach the statistical significance to call a deletion. Thus, the fact that the 44K platform missed this aberration might be explained by tumour heterogeneity and/or the presence of normal DNA in the sample, in addition to the lower probe frequency in comparison to the 180K platform.

The results of this experiment confirmed the ability of the 180K platform to detect more aberrations, especially in heterogeneous tumour DNA. Analysing the WGA test results confirmed the advantage of using the 180K platform. As mentioned earlier, 7 aberrations were detected on the 44K platform using unamplified DNA. The results of the WGA DNA on the 44K platform showed only 2 aberrations; the chr8q amplification and an extra focal amplification on chr7. In contrast, the results for the WGA DNA on the 180K platform were much more consistent with those of the unamplified DNA (Figure 4.17). All of the 12 aberrations detected on the 180K array of the unamplified DNA were detected on the 180K array of the WGA DNA. However, an extra focal deletion on chr17 was noted (Figure 4.17). Therefore, using the 180K platform will allow the use of WGA DNA when required.

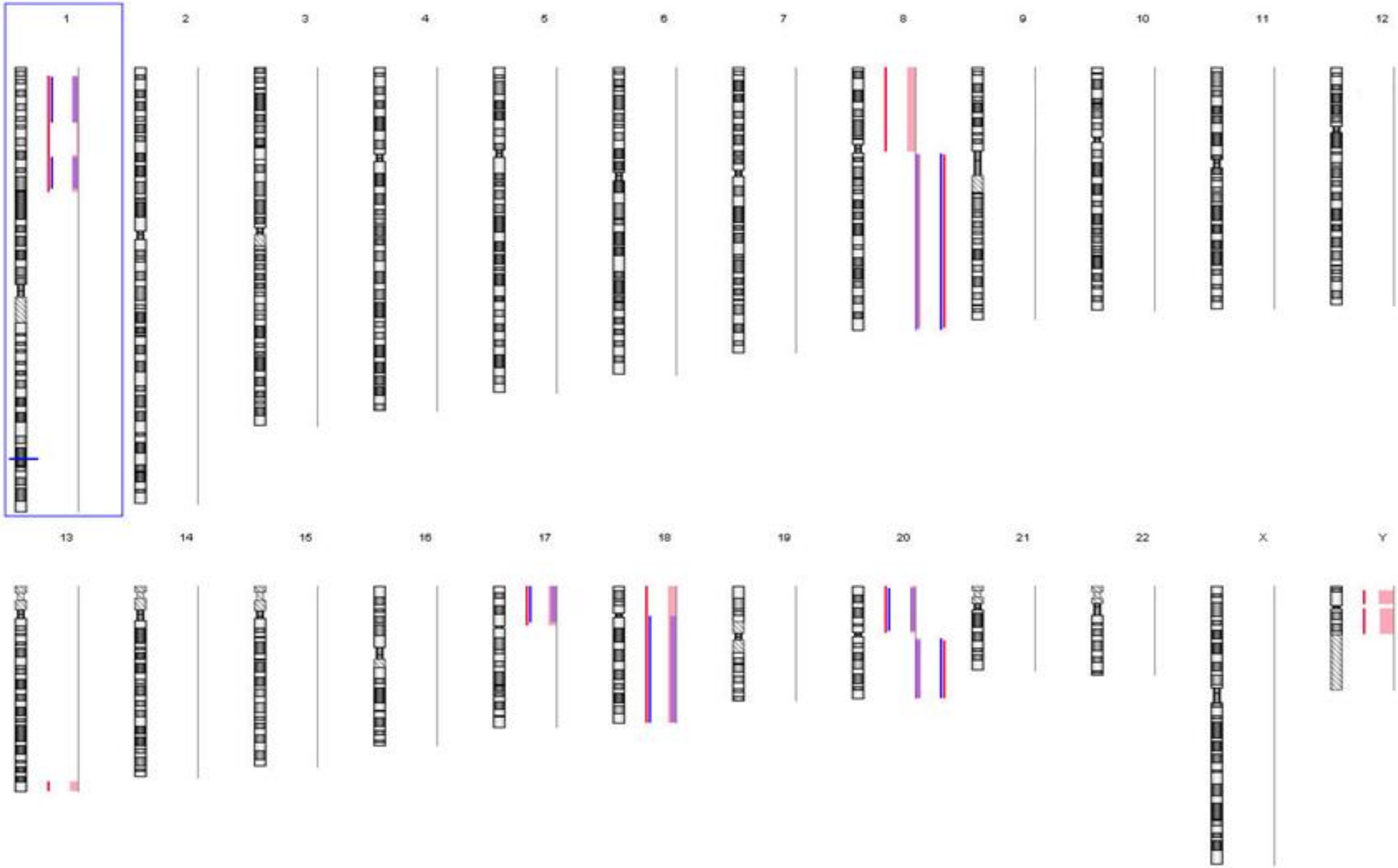


Figure 4.15 CA184 44K vs 180K test

A summary of the genomic aberrations called using both 44K and 180K platforms with the CA184 DNA samples. Solid lines represent aberrations. Transparent lines represent the amplitude of each aberration. Blue and red refer to 44K and 180K respectively.

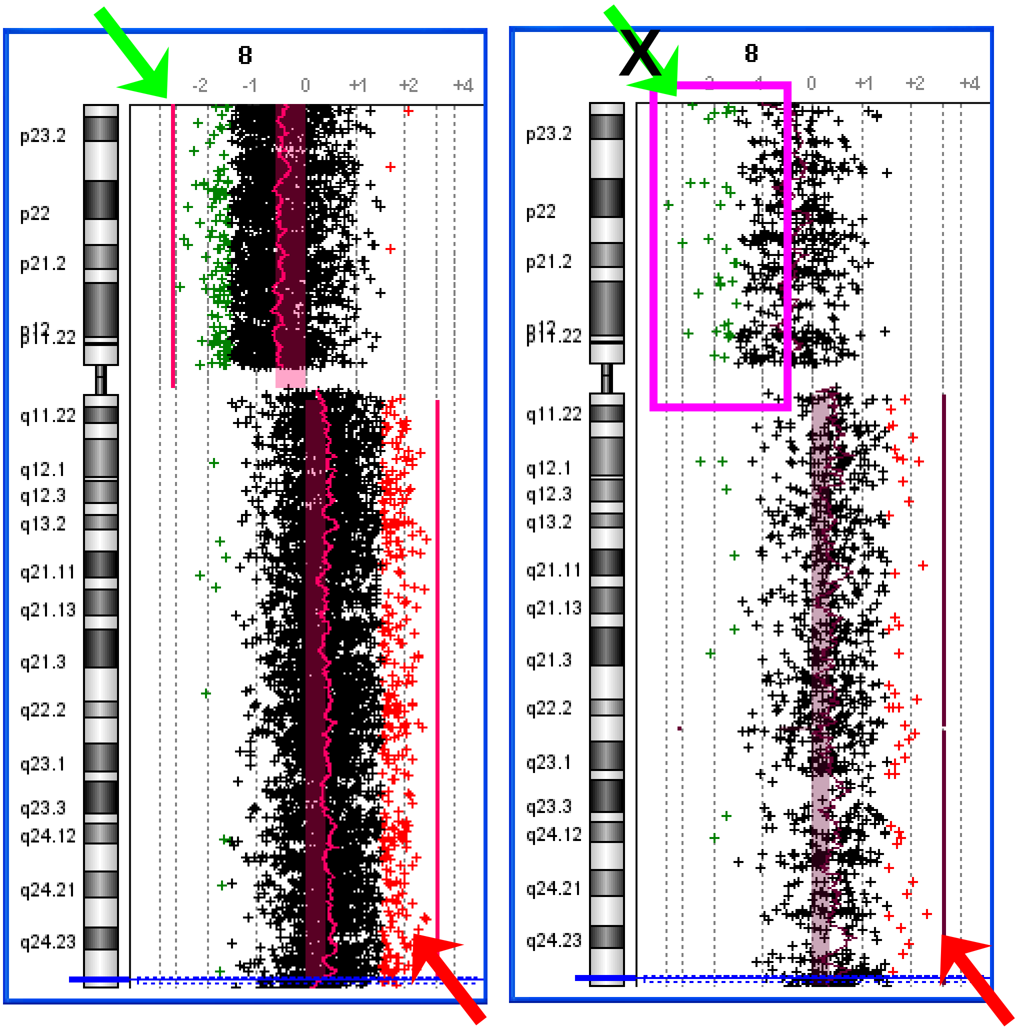


Figure 4.16 CA184 Chromosome 8

A closer look at chromosome 8. Probes (represented by small crosses) coloured in black represent genomic regions with a $\log_2$ratio around zero, probes coloured in red and green represent amplifications ($\log_2$ratio ≥ 0.5) and deletions ($\log_2$ratio ≤ -0.5) respectively. As shown in the Figure, chromosome 8 long arm was called as an amplification on the 2 platforms (marked by red arrows). However, because of the probe density the result appear with higher confidence on the 180K. On the other hand, chromosome 8 short arm was only recognised to be deleted using the 180K format (marked by the green arrow). On the 44K platform, the software did not consider chromosome 8 short arm to be deleted. However, although it is visually clear (marked by pink box), that the area has a deletion, the probe density was not enough for a confident call.

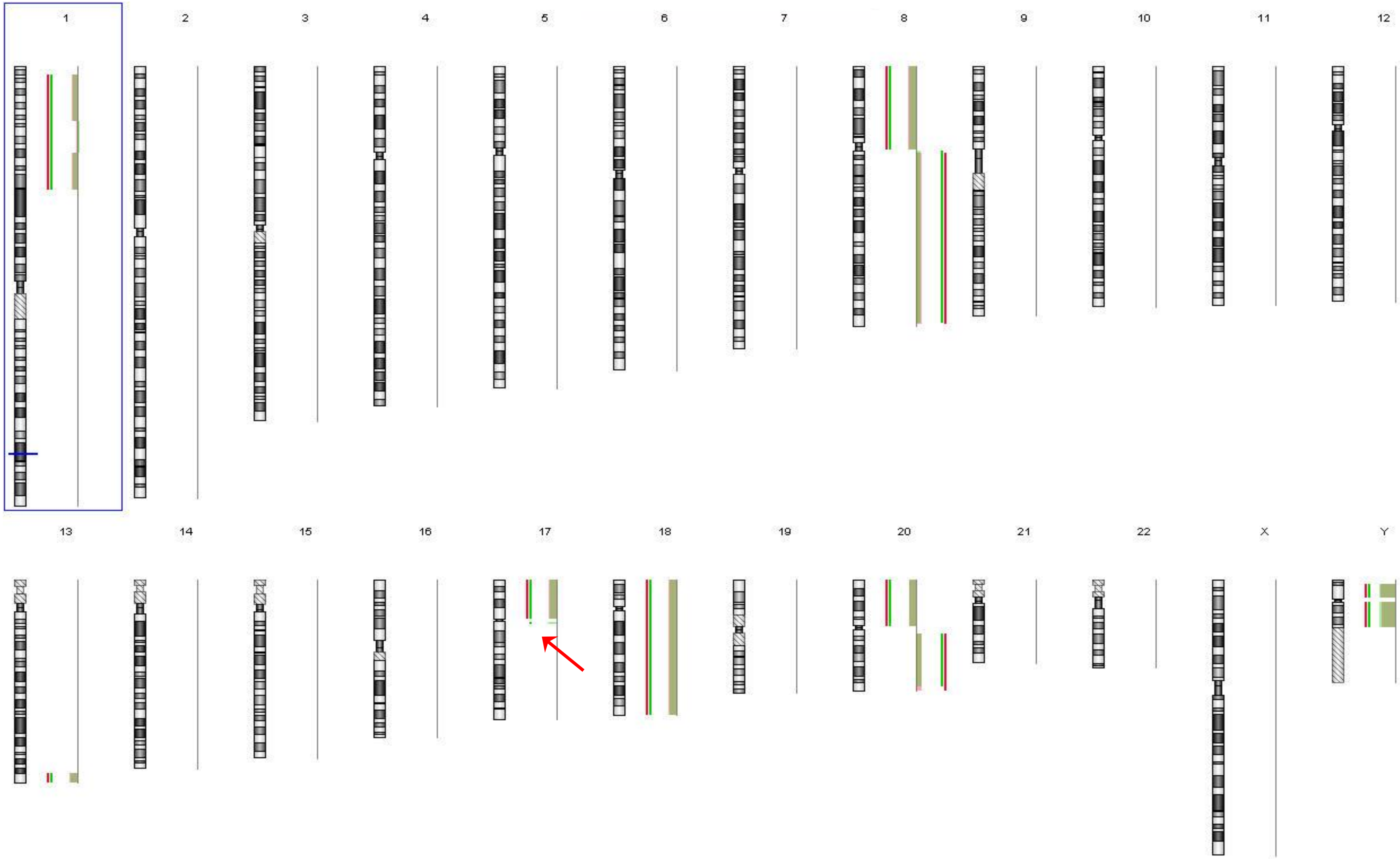


Figure 4.17 CA184 180K platform unamplified vs WGA DNA

A summary of the genomic aberrations called on the 180K platform using the CA184 unamplified and WGA DNA samples. Solid lines represent aberrations. Transparent lines represent the amplitude of each aberration. Red and blue refer to unamplified and WGA respectively. The red arrow points to the extra “spurious” aberration detected in the WGA DNA.

4.2.6.4 aCGH reproducibility test

In order to confirm the specificity and reproducibility of the aCGH technology, 2 extra duplicate experiments were performed. Samples CA080 and CA090 were analysed twice, using both 44K and 180K platforms. The quality of this experiment was very high, since the 4 arrays passed all the QC metrics. The results of the duplicate experiments are shown in Figure 4.18 and Figure 4.19. CA080 had 27 aberrations called using the 44K platform and 38 aberrations called using the 180K platform. All 27 aberrations called using the 44K platform were successfully called based on the 180K platform (Figure 4.18). All the extra aberrations called using the 180K platform were small in size, except the amplification on chr12p and the y-chromosome deletion. However, these were of low amplitude.

Sample CA090 had 39 aberrations called using the 44K platform and 64 using the 180K platform. Out of the 39 called using the 44K array, 38 were also identified with the 180K array (97.4%) (Figure 4.19). All the other aberrations that were called by the 180K, but not the 44K, were relatively small in size except the amplification on chr8q. Figure 4.20 demonstrates an up-close look at one of the large aberrations that was missed by the 44K platform; chr8q from the CA090 duplicate experiment. The probes on the 44K show a shift towards the positive $\log_2$ratio area, however, it was not statistically significant and so not called as an amplification. It is also noteworthy that even on the 180K platform the chr8q amplification was of low amplitude. Figure 4.21 represents a close-up look on a highly concordant result on chr20 from the CA090 experiment. Chr20 had several high amplitude amplifications and deletions of variable sizes and all of them were successfully called on both platforms. The experiments in this section and Section 4.2.6.3 confirm the reproducibility of the aCGH technology and the advantage of using a higher resolution array in detecting small aberrations and those of low amplitude caused by tumour heterogeneity.

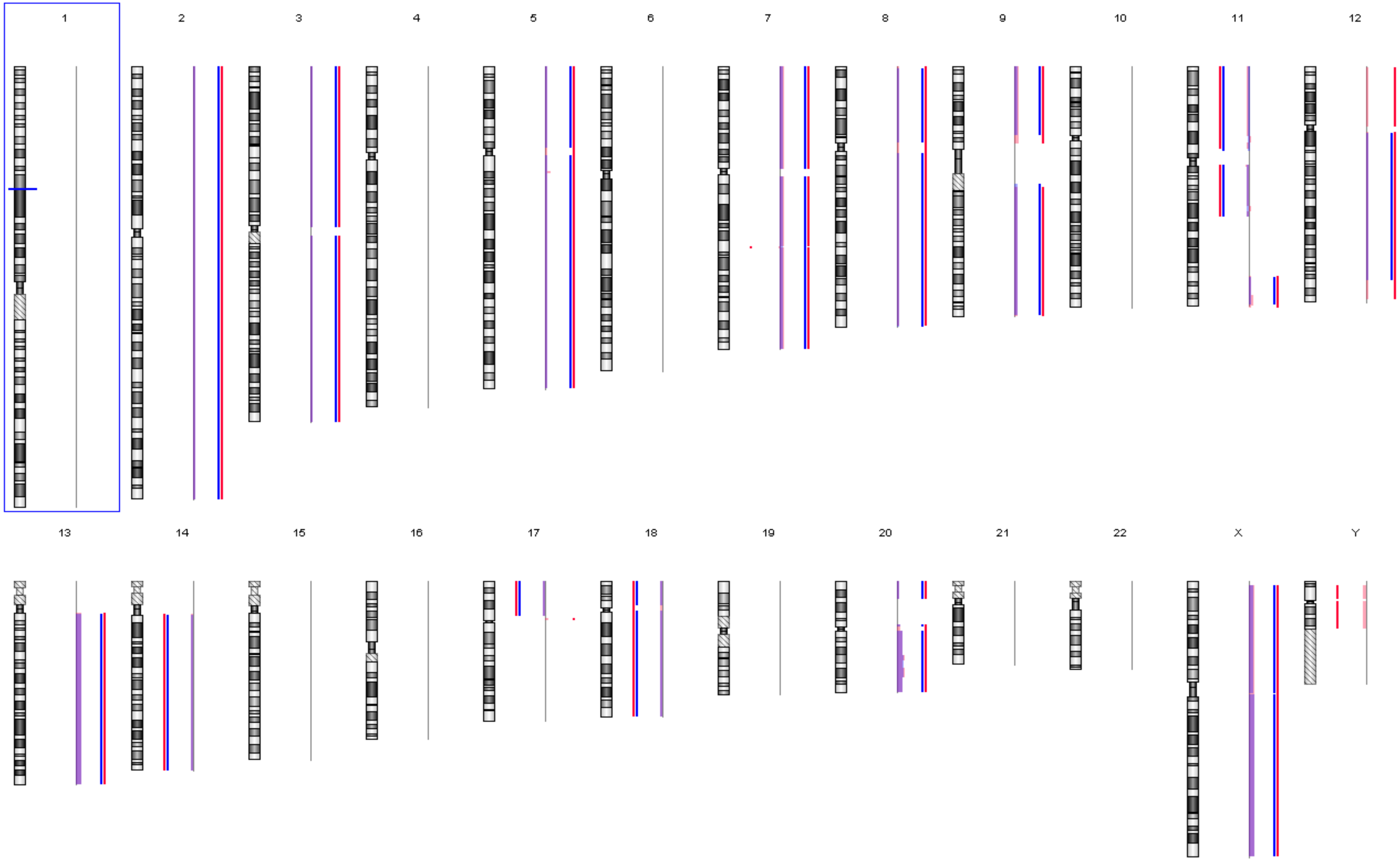


Figure 4.18 CA080 aCGH duplicate experiment

A summary of the genomic aberrations called on both 44K and 180K platforms using the CA080 DNA samples. Solid lines represent aberrations. Transparent lines represent the amplitude of each aberration. Blue and red refer to 44K and 180K respectively.

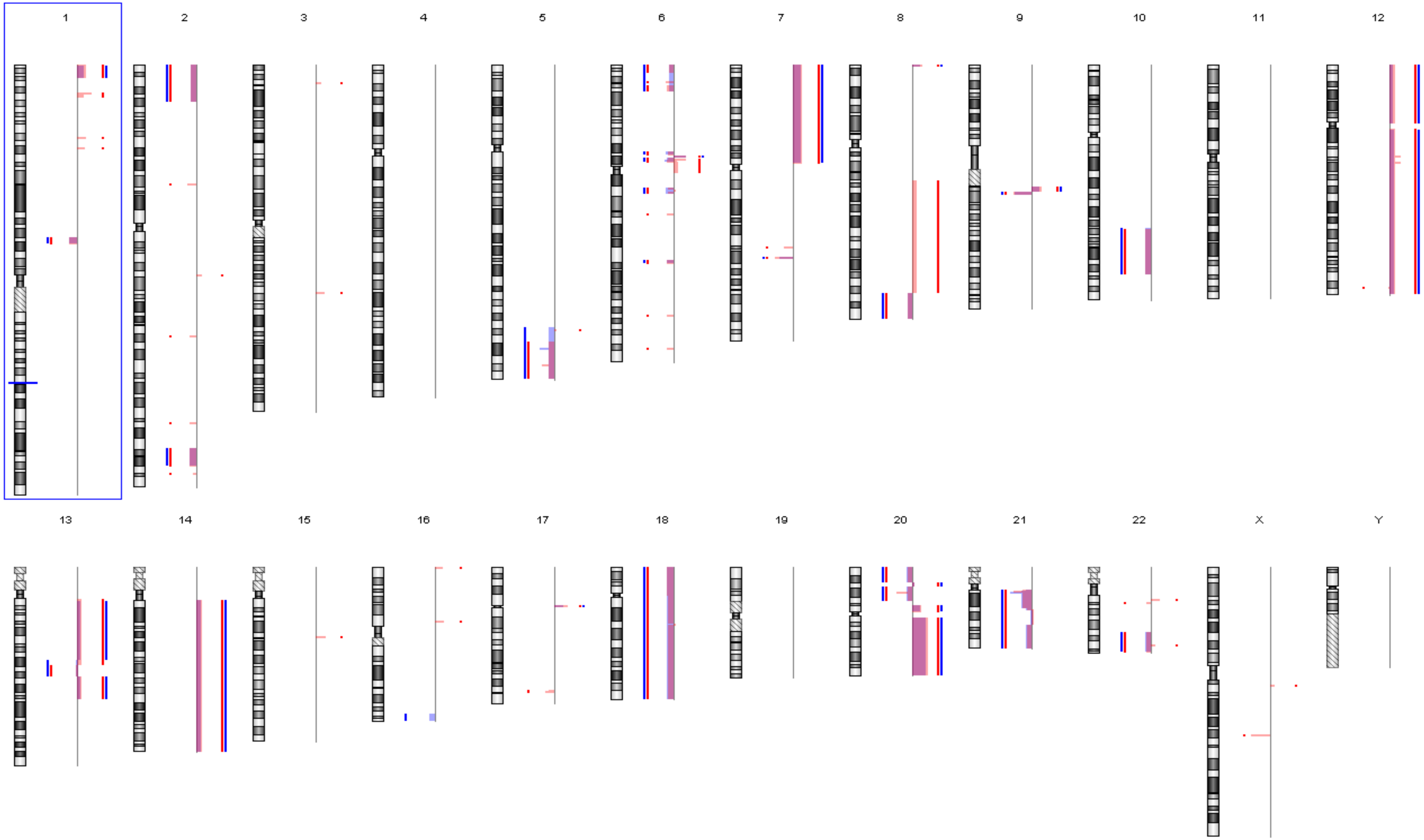


Figure 4.19 CA090 aCGH duplicate experiment

A summary of the genomic aberrations called on both 44K and 180K platforms using the CA090 DNA samples. Solid lines represent aberrations. Transparent lines represent the amplitude of each aberration. Blue and red refer to 44K and 180K respectively.

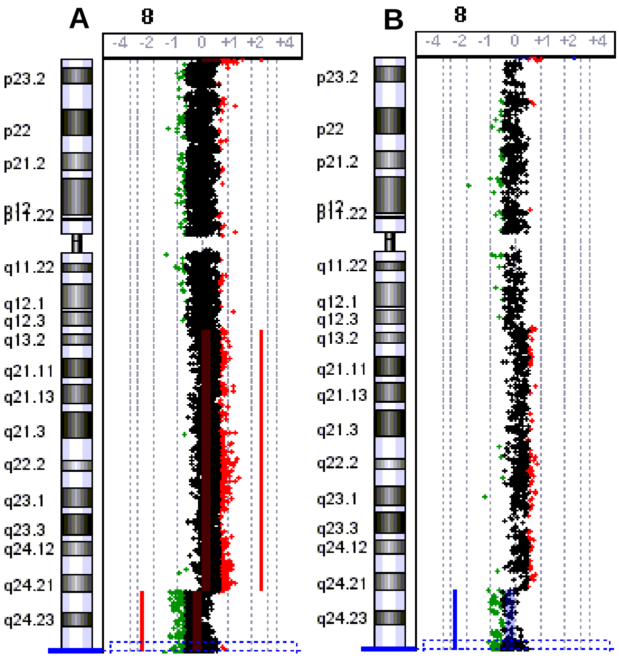


Figure 4.20 CA090 Chromosome 8

Solid lines represent aberrations and transparent boxes represent aberration amplitude. A) CA090 chr8 on the 44K platform, B) CA90 chr8 on the 180K platform, red and green probes represent amplified and deleted regions respectively

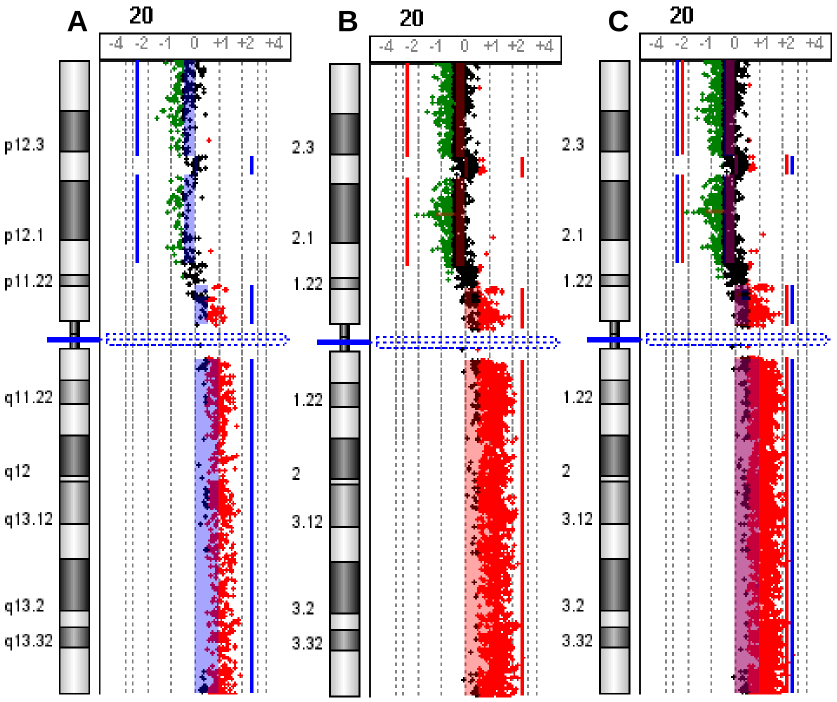


Figure 4.21 CA090 Chromosome 20

Solid lines represent aberrations and transparent boxes represent aberration amplitude. A) CA090 chr20 on the 44K platform B) CA090 chr20 on the 180K platform C) both platforms, red and green probes represent amplified and deleted regions respectively.

4.2.6.5 Chromosomal instability results

So far, there is no clear definition of CIN in CRC. For the purposes of molecular classification as mentioned earlier, CIN was defined as a continuous variable of aberrations and samples were considered chromosomally unstable based on the presence of one or more significant amplifications or deletions in one or more of the chromosomal regions (Cheng *et al.* 2008). Array CGH was successfully performed on the 53 cases and they all passed the QC metrics analysis. Six of the samples were analysed on a 44K platform (CA037, CA112, CA114, CA213, CA214 and CA218) and the remaining 47 samples were analysed on the 180K platform. Samples CA138, CA202, CA203, CA212 and CA249 required WGA prior to their aCGH experiments.

All the samples investigated in the project were analysed using the genomic workbench software v5.0 and the same set of settings (except CA824 and CA1120). Copy number aberrations (CNA) were detected using the ADM2 algorithm with a threshold set to 6.0 and the use of centralisation and fuzzy zero correction. Default feature and aberration filters were applied and intra array probe replicates were combined as described earlier. CA824 and CA1120 were analysed using the same settings, but with ADM1 (instead of ADM2). These 2 arrays had very good QC-metrics (11/11), however, ADM2 failed to call any CNA for unapparent reasons. Visual inspection of the aCGH data has shown clear CNA which were successfully called by ADM1. Agilent bioinformatics support recommended the use of ADM1 for these 2 samples. Out of the 53 cases, 5 (CA023, CA037, CA046, CA110, and CA248) were considered to be be chromosomally stable (i.e. no CNA). The rest (n=48, 91%) were considered chromosomally unstable. Table 4.12 summarises the number of aberrations (deletions and amplifications) for all of the chromosomally unstable samples.

Samples CA824, CA1120 and CA184 aCGH data was used only to determine their overall CIN status (all chromosomally unstable). However, they were excluded from any detailed chromosomal aberration analysis. CA824 and CA1120 were excluded because they were analysed using ADM1 and CA184 was excluded for having high DLRS ratio (Section 2.2.15.10).

Table 4.12 Aberrations summary per sample

Sample	Aberrations	Amp	Del	CIMP	*TP53*	Gender	Survival/years	Dukes Stage	Tumour Location
CA112	1	1	0	CIMP-L	WT	Female	2.92	C1	Distal
CA795	2	2	0	CIMP-L	WT	Female	NA	C	Proximal
CA1120	2	1	1	CIMP-L	Mt	Male	NA	C	Proximal
CA086	3	2	1	CIMP-L	Mt	Female	0.92	C1	Distal
CA016	4	3	1	CIMP-L	WT	Female	NA	B	Proximal
CA150	5	0	5	CIMP-L	WT	Male	NA	C	Distal
CA83	10	9	1	CIMP-L	WT	Male	NA	Unknown	Distal
CA184	12	2	10	CIMP-L	Mt	Male	6.50	C1	Distal
CA135	14	3	11	CIMP-L	WT	Female	2.83	B	Distal
CA167	14	6	8	CIMP-L	WT	Male	Alive	C1	Distal
CA218	14	8	6	CIMP-L	Mt	Female	Alive	C1	Distal
CA109	18	10	8	CIMP-H	WT	Male	1.00	C2	Proximal
CA122	18	17	1	CIMP-N	WT	Female	NA	C	Proximal
CA632	18	2	16	CIMP-L	Mt	Female	NA	C	Proximal
CA824	21	12	9	CIMP-L	WT	Female	NA	C	Distal
CA201	23	14	9	CIMP-L	WT	Female	2.50	B	Distal
CA138	24	8	16	CIMP-H	WT	Male	0.58	C1	Distal
CA212	29	0	29	CIMP-N	WT	Female	NA	C1	Proximal
CA202	33	25	8	CIMP-L	WT	Male	Alive	A	Distal
CA213	33	7	26	CIMP-L	Mt	Female	Alive	C2	Distal
CA208	35	19	16	CIMP-L	Mt	Female	0.08	C1	Distal
CA104	37	17	20	CIMP-L	Mt	Female	0.58	C2	Distal
CA080	38	27	11	CIMP-L	WT	Male	Alive	B	Distal
CA221	39	17	22	CIMP-N	WT	Male	0.75	B	Distal
CA085	43	20	23	CIMP-L	Mt	Male	NA	B	Proximal
CA097	43	17	26	CIMP-L	Mt	Male	3.33	Unknown	Distal
CA088	45	13	32	CIMP-H	Mt	Female	0.83	C1	Proximal
CA206	50	34	16	CIMP-L	Mt	Female	3.25	B	Distal
CA249	61	33	28	CIMP-L	Mt	Male	7.25	B	Distal
CA090	63	26	37	CIMP-L	Mt	Female	7.58	B	Distal
CA271	66	47	19	CIMP-H	Mt	Male	Alive	C1	Proximal
CA741	66	35	31	CIMP-L	WT	Female	NA	C	Proximal
CA203	67	21	46	CIMP-L	WT	Female	5.75	B	Proximal
CA079	71	19	52	CIMP-N	Mt	Male	0.42	C1	Distal
CA114	74	13	61	CIMP-H	WT	Male	2.58	C1	Distal
CA223	76	53	23	CIMP-H	Mt	Male	0.67	C2	Proximal
CA214	86	12	74	CIMP-L	WT	Male	Alive	B	Distal
CA158	87	31	56	CIMP-N	Mt	Female	NA	Unknown	Proximal
CA053	88	49	39	CIMP-L	Mt	Female	Alive	B	Distal
CA098	93	38	55	CIMP-L	Mt	Male	4.75	B	Distal
CA1350	94	85	9	CIMP-L	WT	Female	NA	B	Distal
CA107	110	91	19	CIMP-L	WT	Male	1.58	D	Distal
CA863	114	61	53	CIMP-L	Mt	Male	NA	A	Proximal
CA153	150	100	50	CIMP-L	Mt	Male	NA	C	Proximal
CA142	219	171	48	CIMP-L	Mt	Female	4.75	C1	Proximal
CA828	248	152	96	CIMP-L	Mt	Male	NA	A	Distal
CA45	260	120	140	CIMP-L	Mt	Female	5.75	C1	Distal
CA244	411	116	295	CIMP-L	Mt	Male	Alive	C1	Distal

Table 4.12 summarises the number of aberrations/ sample including the type of the aberration (Amplification or deletion). *TP53* mutation status, gender, survival years and tumour stage are also summarised.

4.2.6.5.1 Copy number aberrations: summary and patterns

A total of 3097 CNA were identified in the 45 chromosomally unstable samples (apart from CA824 and CA184) (Table 4.12). The number of CNA per sample ranged from 1-411, with an average of 68.8 per sample. CNA ranged in size from 0.0136Mb-147.4767Mb (mean: 14.7627Mb, median: 2.2910Mb). Out of the 3097 CNA identified, 1554 were amplifications, ranging in size from 0.0136Mb-147.4767Mb (mean: 17.5479Mb, median: 3.2727Mb). There were also 1543 deletions, ranging in size from 0.0245Mb-145.1558 (mean: 11.9577Mb, median: 1.5526Mb). An overview of the pattern and frequencies of these CNA is presented in Figure 4.22. A summary of the CNA identified in the 40 samples analysed using the 180K arrays is presented in Figure 4.23. In the 45 CIN cases, the most common amplifications were within chr 20q (73.3%, n=33), 13 (57.8%, n=26), 8q (53.3%, n=24), 7 (51.1%, n=23) and X (51.1%, n=23) and the most common deletions were within chr 18 (55.6%, n=25), 8p (51.1%, n=23) and 17p (51.1%, n=23). Some of these regions contain key CRC driver genes, such as *MYC* at 8q, *SMAD4* at 18q and *TP53* at 17p.

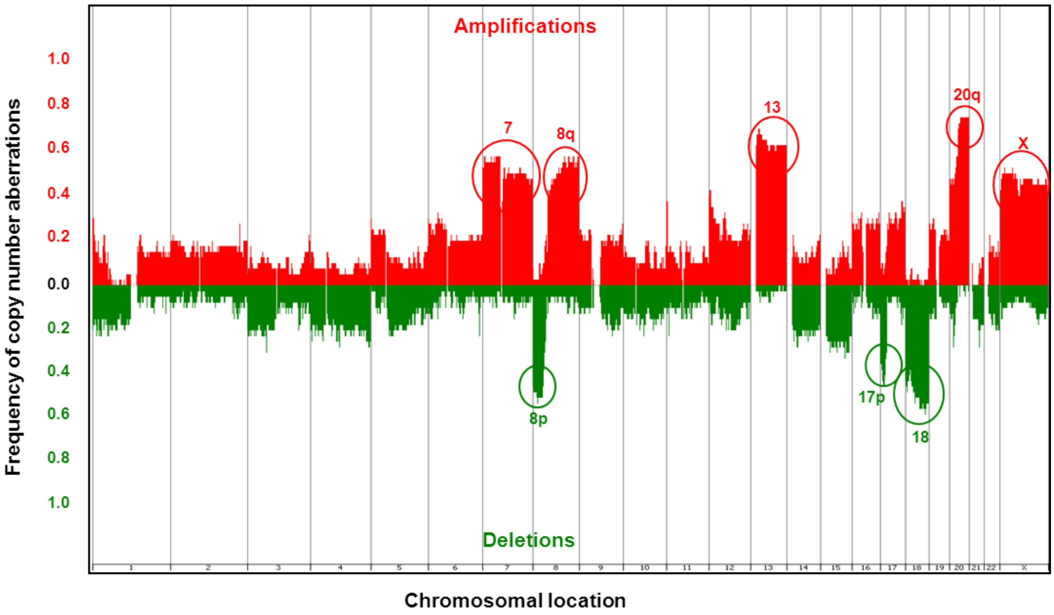


Figure 4.22 Patterns of CNA in the studied cohort.

Red represents amplifications and green represents deletions. The y-axis of the red and green bars reflects frequency. The figure represents the 40 CIN cases analysed by the 180K format.

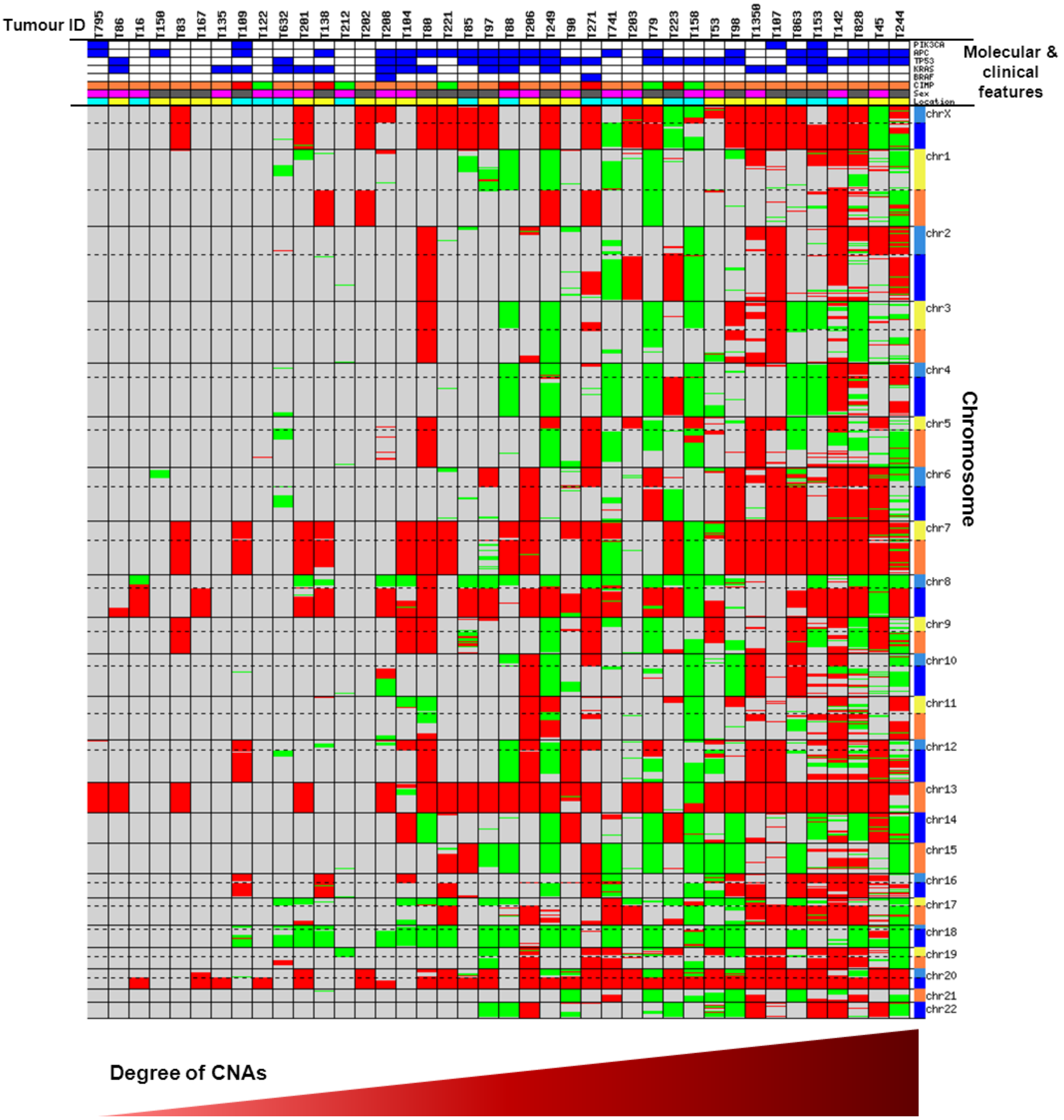


Figure 4.23 CNA identified in 40 chromosomally unstable cases (180K)

Representation of the entire CNA identified in 40 chromosomally unstable cases analysed using the 180K platform. Red represents amplification and green represents deletion. Dotted lines mark centromeres. Molecular and clinical features in order from top; *PIK3CA*, *APC*, *TP53*, *KRAS* and *BRAF* mutation status (blue and white represents mutant and WT respectively), CIMP (red, orange and green represent CIMP-H, CIMP-L and CIMP-N respectively), patient’s sex (pink and grey represent female and male respectively) and tumour location (yellow and cyan represent proximal and distal respectively).

4.2.7 Associations of molecular events

4.2.7.1 CIMP status and molecular and clinical features

Two-tailed Fisher's exact test was used to test whether the published associations between the different molecular events and tumour/patient characteristics could be seen in our data. A summary of the molecular and clinical characteristics of the cases in relation to their CIMP status is summarised in Table 4.13. CIMP-H was found to associate with mutations in *KRAS* and *BRAF* in comparison to CIMP-L/N (p-value = 0.024) and the only sample with the *BRAF* V600E mutation was CIMP-H (the only sample with methylation in all of the markers). However, no association was found between CIMP status and *PIK3CA* mutations (p-value = 0.5323), patient's gender (p-value = 0.1959) and tumour location (p-value = 0.6687). Moreover, none of the CIMP-N (n=6) cases had any mutations in *KRAS*, *BRAF* or *PIK3CA*. Compared to CIMP-H/N, CIMP-L tumours seemed to be more common among samples from the distal colon, but the results did not reach statistical significance (p-value = 0.09). Figure 4.24 represents the CIMP status in relation to the molecular and clinical features of the 53 patients.

Table 4.13 CIMP status and molecular and clinical features

		CIMP status		
		n (%)		
		CIMP-H	**CIMP-L**	**CIMP-N**
		6 (100.0)	**41 (100.0)**	**6 (100.0)**
KRAS/BRAF	Mutant	5 (83.3)	15 (36.6)	0 (0.0)
	WT	1 (16.7)	26 (63.4)	6 (100.0)
PIK3CA	Mutant	1 (16.7)	5 (12.2)	0 (0.0)
	WT	5 (83.3)	36 (87.8)	6 (100.0)
TP53	Mutant	3 (50.0)	21 (51.2)	2 (33.3)
	WT	3 (50.0)	20 (48.8)	4 (66.7)
APC	Mutant	4 (66.7)	21 (51.2)	2 (33.3)
	WT	2 (33.3)	20 (48.8)	4 (66.7)
Tumour Location	Proximal	4 (66.7)	12 (29.3)	3 (50.0)
	Distal	2 (33.3)	29 (70.7)	3 (50.0)
Gender	Male	5 (83.3)	20 (48.8)	3 (50.0)
	Female	1 (16.7)	21 (51.2)	3 (50.0)
CIN	CIN+	6 (100.0)	37 (90.2)	5 (83.3)
	CIN-	0 (0.0)	4 (9.8)	1 (16.7)

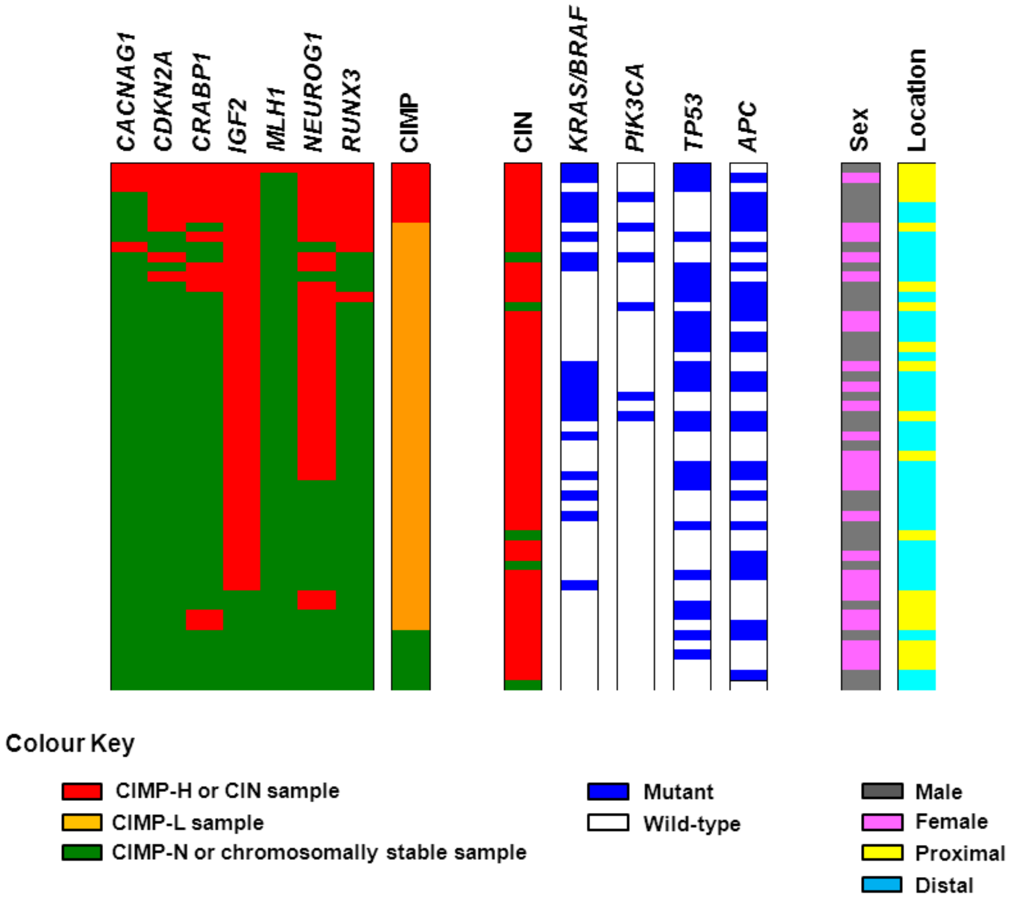


Figure 4.24 CIMP and molecular and clinical features

4.2.7.2 CIMP and CIN

CIMP-H samples were previously shown to be inversely correlated with CIN (Goel *et al.* 2007) or correlated with a low degree of CIN when compared with CIMP-L/N cases (Cheng *et al.* 2008). Table 4.14 summarises the range of chromosomal aberrations in CIMP-H and CIMP-L/N samples. CIMP-H and CIMP-L/N were not found to be significantly different in-terms of the number of chromosomal aberrations (Mann Whitney U-test, p-value = 0.88) (Figure 4.25).

Although the overall CNA frequency was not significantly different between CIMP-H and CIMP-L/N, there were some differences in the frequency of some of the broad and common CNA. The main differences were observed in amplifications within 13q

(33.3% CIMP-H, 61.5% CIMP-L/N), 20q (33.3% CIMP-H, 79.5% CIMP-L/N), chrX (16.7% CIMP-H, 56.4% CIMP-L/N) and chr7 (83.3% CIMP-H, 46.2% CIMP-L/N) and deletions in 8p (83.3% CIMP-H, 46.2% CIMP-L/N). However, both CIMP-H and CIMP-L/N tumours exhibited similar frequencies of amplifications within 8q (66.7% CIMP-H, 51.3 CIMP-L/N), and deletions in chr18 (66.7% CIMP-H, 53.8% CIMP-L/N) and 17p (66.7% CIMP-H and 48.7% CIMP-L/N).

Table 4.14 CIN and CIMP

	Copy number aberrations (CIN)			
	No. of samples	**No. of Aberrations (range)**	**Mean**	**Median**
CIMP-H	6	18-76	50.50	55.50
CIMP-L/N	39	1-411	71.64	43.00

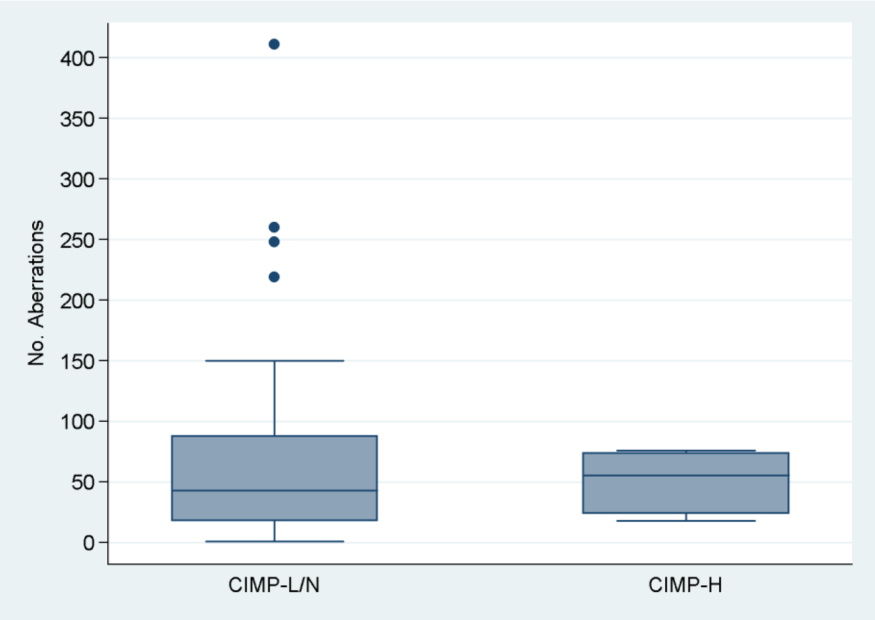


Figure 4.25 CIMP status and chromosomal aberrations

Box plot and whisker charts representing the range and mean of chromosomal aberrations in CIMP-L/N cases compared to CIMP-H cases.

4.2.7.3 CIN and molecular and clinical features

A summary of the molecular and clinical characteristics of the 53 cases in relation to CIN status is presented in Table 4.15. Chromosomally stable samples did not have any *TP53* mutations while CIN samples had high frequency (54.2%), however, this difference was borderline statistically significant (p-value = 0.051). Moreover, CIN samples had relatively lower *PIK3CA* mutation frequency, but this was not statistically significant (p-value = 0.093). No associations were found with patient's gender (p-value = 0.356), or tumour location (p-value = 0.613) or mutations in *KRAS* (p-value = 1.000), *BRAF* (p-value = 1.000), or *APC* (p-value = 0.669).

Table 4.15 CIN status and molecular and clinical features

		CIN status		
		n (%)		
		CIN positive	CIN negative	
		48 (100.0)	5 (100.0)	p-value
KRAS	Mutant	17 (35.4)	2 (40.0)	1.000
	WT	31 (64.6)	3 (60.0)	
BRAF	Mutant	2 (4.2)	0 (0.0)	1.000
	WT	46 (95.8)	5 (100.0)	
PIK3CA	Mutant	4 (8.3)	2 (40.0)	0.093
	WT	44 (91.7)	3 (60.0)	
TP53	Mutant	26 (54.2)	0 (0.0)	0.051
	WT	22 (45.8)	5 (100.0)	
APC	Mutant	25 (52.1)	2 (40.0)	0.669
	WT	23 (47.9)	3 (60.0)	
Tumour Location	Proximal	17 (35.4)	2 (40.0)	1.000
	Distal	31 (64.6)	3 (60.0)	
Gender	Male	24 (50.0)	4 (80.0)	0.355
	Female	24 (50.0)	1 (20.0)	

4.2.8 Molecular classification of the tested samples

All the tested samples were MSS, therefore, none of them will fit within the MSI-H pathway. Out of the 5 chromosomally stable samples, 4 were CIMP-L. Table 4.14 summarises the molecular characteristics of the 4 chromosomally stable/CIMP-L samples. These samples can fit within the heterogeneous CIMP-L subtype described in Section 1.3.5.1.3.

Table 4.16 Molecular and clinical features of the CIN negative samples

Sample ID	*BRAF*	*KRAS*	*TP53*	*APC*	*PIK3CA*	Location	Gender	Age
CA023	WT	WT	WT	Mut	Mut	Proximal	Male	79
CA037	WT	Mut	WT	Mut	WT	Distal	Male	57
CA046	WT	Mut	WT	WT	Mut	Distal	Female	65
CA110	WT	WT	WT	WT	WT	Proximal	Male	78

One of the chromosomally stable samples (CA248) was also CIMP-N (and MSS). Therefore, CA248 fits within the chromosomally stable, CIMP-N, MSS subtype described in Section 1.3.5.1. Sample CA248 is referred to in the following section. The remainder of the samples were identified as chromosomally unstable. Six of these samples were also CIMP-H. The 6 CIMP-H samples did not differ from the rest of the CIN samples in terms of frequency of aberrations. A summary of the molecular and clinical features of all the CIN positive samples stratified according to their CIMP status (CIMP-H and CIMP-L/N) is presented in Table 4.17. Two tailed Fisher’s exact test was used to test for differences between the 2 groups (CIN/CIMP-H and CIN/CIMP-L/N). Neither groups had any significant difference in terms of tumour location (p-value = 0.167), patient’s gender (p-value = 0.188), or mutations in *PIK3CA* (p-value = 0.425), *TP53* (p-value = 1.000) and *APC* (p-value = 0.668). However, the CIN/CIMP-L/N had significantly fewer *KRAS*/*BRAF* mutations (p-value = 0.022). In summary, the analysed cohort can be divided into 3 molecular subgroups; chromosomally stable CIMP-L/MSS (11.9%), CIN/CIMP-H/MSS (14.3%) and CIN/MSS subgroup (71.6%), in addition to CA248 which did not show any form of genomic instability. Figure 4.26 represents the 3 molecular subtypes identified in the screened samples.

Table 4.17 CIN positive samples and CIMP status

CIN positive samples		CIMP status n (%) CIMP-H 6 (100.0)	 CIMP-L/N 42 (100.0)	p-value
KRAS/BRAF	Mutant	5 (83.3)	13 (31.0)	0.022
	WT	1 (16.7)	29 (69.0)	
PIK3CA	Mutant	1 (16.7)	3 (7.1)	0.425
	WT	5 (83.3)	39 (92.9)	
TP53	Mutant	3 (50.0)	23 (54.8)	1.000
	WT	3 (50.0)	19 (45.2)	
APC	Mutant	4 (66.7)	21 (50.0)	0.668
	WT	2 (33.3)	21 (50.0)	
Tumour Location	Proximal	4 (66.7)	13 (31.0)	0.167
	Distal	2 (33.3)	29 (69.0)	
Gender	Male	5 (83.3)	19 (45.2)	0.188
	Female	1 (16.7)	23 (54.8)	

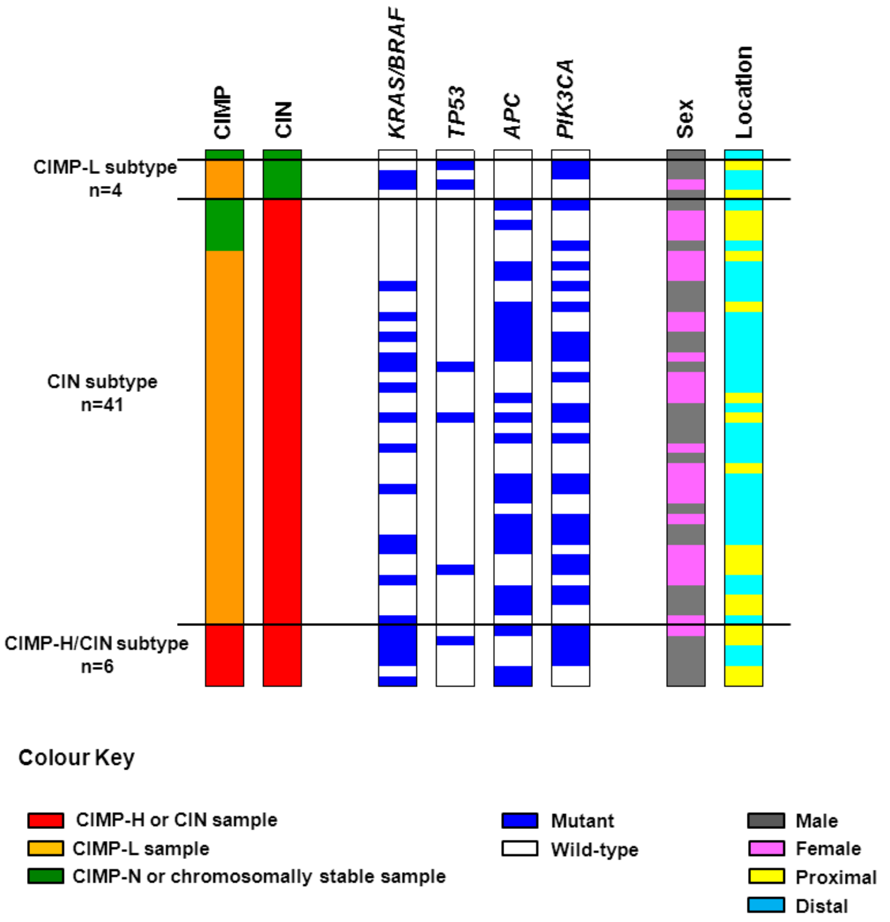


Figure 4.26 The molecular subtypes identified in the screened cohort

* Mutations details, in addition to CIMP and CIN are summarised in Appendix 6.

4.2.9 Molecular characteristics & CRC phenotype (extreme cases)

One of the investigated samples (CA248) was shown to be MSS, chromosomally stable and with CIMP-N. This subset of CRC was previously reported at a similar low frequency (Cheng *et al.* 2008). However, surprisingly, the patient (CA248) did not even have any of the common somatic key mutations discussed above. The patient was diagnosed with CRC in the year 2002 at the age of 65 without any family history of cancer. The cancer had very good prognosis, it responded well to treatment, and cancer registry records of March 2011 show that the patient was still alive at this time.

On the other hand, one of the cases (CA208) had 5 somatic mutations; *APC*, *BRAF* and *KRAS*, in addition to 2 somatic mutations in *TP53*. Moreover, the tumour was chromosomally unstable with CIMP low (but MSS). The patient did not have any family history of cancer, but she was diagnosed with CRC at the age of 51 and died within 1 month of diagnosis due to aggressive cancer progression. The aggressive CRC behaviour and the bad prognosis might be explained by the complex molecular characteristics of the tumour.

4.3 Discussion

As described earlier, molecular classification of CRC can play an important role in research and in the clinic. Towards the beginning of the project, 4 main studies were published on the molecular classification of CRC (Chang *et al.* 2006, Shen *et al.* 2007, Cheng *et al.* 2008, Derks *et al.* 2008). One of the drawbacks of these initial studies was the lack of standardisation, especially when analysing CIN and CIMP, which makes it very difficult to compare the findings. The initial aim of this project was to try and classify our CRC samples using more standardised methods and marker panels.

4.3.1 Quality measures and MSI testing

A thorough molecular characterisation was performed on our samples. The matched normal/tumour DNA samples were verified as coming from the same patient. MSI analysis was the first step of our classification, and a relatively new marker panel was used. This panel was developed according to the revised NCI recommendations on MSI testing (Bacher *et al.* 2004). Two samples out of 57 were shown to be MSI-H, however, they were excluded from further study because of contamination with a different DNA sample in sample CA945, and because of high levels of normal DNA in sample CA008. The remaining samples were all considered to be MSS. Because of the size of our sample, the expected number of MSI-H cases was ~6-9 based on a known MSI frequency of ~10-15% of sporadic CRC. The frequency of MSI-H in our samples (n=2, 3.6%) was below the expected. This could be due to sampling variation (i.e. chance), or due to limitations in the detection technique. We determined the sensitivity of the MSI detection kit by testing a serial dilution of MSI-H DNA extracted from HCT-116 cell line DNA. HCT-116 cells are classified as the MSI-H/CIMP-H subtype. The analysis showed the ability of the kit to detect MSI down to 10% dilution (assuming careful examination). Logically, therefore, only MSI-H tumours with a very high level of normal DNA contamination would have been missed by this assay. Thus, we considered the remaining cases as MSS.

4.3.2 Sporadic CRC driver mutations

The paired cohort DNA samples were tested for the most common mutations of the key CRC driver genes; *APC*, *TP53*, *KRAS*, *BRAF* and *PIK3CA*. The frequencies of the mutations found, especially *BRAF* (3.8%) and *PIK3CA* (11.3%), are consistent with the MSS nature of our cohort. Mutations in both genes were previously shown to cluster with MSI-H samples and the frequencies of mutations in both genes in our cohort were below those expected for a sporadic CRC cohort (Weisenberger *et al.* 2006, Abubaker *et al.* 2008). In contrast, the frequency and pattern of *APC* (54.7%), *TP53* (49%) and *KRAS* (36%) mutations were within the expected range for sporadic CRC. These mutations are more common in sporadic CRC than *BRAF* and P*IK3CA* and they have a wider range of mutations.

4.3.3 CpG island methylator phenotype and MS-MLPA

The samples were screened for CIMP using a set of 7 methylation markers (*CACNA1G*, *CDKN2A*, *CRABP1*, *IGF2*, *MLH1*, *NEUROG1* and *RUNX3*), which had been shown in a previous large study to specifically and accurately estimate the genomic CIMP status. MS-MLPA was used to estimate CIMP (Ogino *et al.* 2007). None of the studies in CRC have previously used MS-MLPA and the kit used (ME042-A1) was still under-development by MRC-Holland. The performance and accuracy of the kit was tested and the results confirm the ability of the MS-MLPA kit to accurately estimate methylation levels. In summary, out of the 53 tested samples, 6 (11.3%) were defined as CIMP-H, 41 (77.4%) as CIMP-L and 6 (11.3%) were defined as CIMP-N. The CIMP-H frequency was less than expected, (~20%), for sporadic CRC (Cheng *et al.* 2008). However, this can be explained by the MSS nature of the analysed cohort. CIMP-H and MSI-H are known to strongly associate together in sporadic CRC (Weisenberger *et al.* 2006). The absence of *KRAS*, *BRAF* and *PIK3CA* mutations from the CIMP-N subgroup confirms the existence of CIMP-L as an independent subgroup of CIMP with a distinct biological background (Ogino and Goel 2008).

4.3.4 Chromosomal instability and array CGH

Finally, CIN was investigated in the samples using aCGH. High-resolution genome wide aCGH platforms were used and their intrinsic reproducibility was confirmed. Our samples exhibit a high frequency of CIN (90.1%, n=48), which is higher than expected for sporadic CRC samples (65-85%) (Pino and Chung 2010, Migliore *et al.* 2011). CIN samples had a higher frequency of *TP53* mutations (p-value = 0.051) and none of the chromosomally stable samples had mutations in *TP53* which agree with previously published data (Chang *et al.* 2006).

The overall pattern of identified broad CNA, amplifications within chromosomes 20q (73.3%, n=33), 13 (57.8%, n=26), 8q (53.3%, n=24), 7 (51.1%, n=23) and X (51.1%, n=23) and deletions within chromosomes 18 (55.6%, n=25), 8p (51.1%, n=23) and 17p (51.1%, n=23), strongly agree with published data (Hermsen *et al.* 2002, Lassmann *et al.* 2007, Poulogiannis *et al.* 2010a)

4.3.5 Molecular associations and CRC subtypes

In summary, 3 CRC molecular subtypes were observed in our cohort (Figure 4.26). Chromosomally stable/CIMP-L characterised by lower levels of CIMP, MSS and absence of chromosomal aberrations and *TP53* mutations. The 2nd subtype was the CIN subtype characterised by CIN, MSS, *TP53* mutations and distal tumours. The 3rd subtype was characterised by high levels of CIMP, CIN, MSS and *KRAS* and *BRAF* mutations. These results clearly confirm the molecular heterogeneity of sporadic CRC. Statistical analyses have also confirmed previous correlations between CIMP-H and *KRAS/BRAF* mutations. Finally, the known association between *TP53* mutations and CIN was also confirmed.

Interestingly, our CIMP-H samples were neither mutually exclusive with CIN nor correlated with low degree of CIN, as previously reported (Goel *et al.* 2007, Cheng *et al.* 2008). CIMP-H samples were characterised with CIN and they were not statistically different from other CIN samples (CIMP-L/N) (Mann Whitney U-test, p-value = 0.88). This could be the result of sample size, however, the previous study that has shown a mutual exclusive relationship between CIMP-H and CIN employed LOH analysis to define CIN (Goel *et al.* 2007). LOH analysis is known to underestimate CIN and result in false negative chromosomally stable cases (Ogino and Goel 2008). On the other hand, the cases in the study that established the association between CIMP-H and lower degree of chromosomal aberrations were mainly CIMP-H/MSI-H cases (Cheng *et al.* 2008). MSI and CIN have a well-established mutually exclusive relationship (Ogino and Goel 2008). Therefore, the association between CIMP-H and the lower degree of CIN in this study could well be MSI dependent. Close examination of the data from the study by Cheng and colleagues has confirmed this suggestion. Four of the 60 investigated cases (6.67%) were CIMP-H/MSS samples and they had more chromosomal aberrations than CIMP-H/MSI-H samples. In fact, one case with the highest number of chromosomal aberrations was a CIMP-H/MSS sample. In their conclusion, the authors did suggest that the association between CIMP-H and low degree of CIN could reflect the inverse relationship between CIN and MSI and they anticipated the possible existence of the CIN/CIMP-H subset (Cheng *et al.* 2008). Nevertheless, a CIN/CIMP-H/MSS subtype has not been proposed by any of the current CRC molecular classification systems (Jass 2007, Issa 2008, Ogino and Goel 2008). Based on our data and the suggestion made by Cheng and colleagues, we propose the existence of a novel CIMP-H/MSS CRC subtype with high levels of CIN. However, this subtype needs to be confirmed in a larger sample size with more CIMP-H/MSS samples.

Although the frequency of the CNA was similar across chromosomally unstable samples regardless of CIMP status, there were some differences in terms of common broad CNA patterns. Amplifications within 13q, 20q and chrX were less frequent in the CIMP-H samples. However, chr7 amplifications and 8p deletions were less common in the CIMP-L/N samples. Similar differences in the frequencies of amplifications within 13q and 20q were recently observed between CIN/MSS tumours and the rare CIN/MSI-H tumours (Poulogiannis *et al.* 2010a). These observations indicate that 13q and 20q amplifications play an important and specific role in the samples developing primarily through the CIN pathway. In contrast, CIMP-H and

CIMP-L/N chromosomally unstable samples had similar frequencies of 8q amplifications and 18q and 17p deletions. Likewise, MSI-H and MSS chromosomally unstable samples showed the same pattern for 8q amplifications (Poulogiannis *et al.* 2010a). This might indicate that 8q aberrations might play a role in the development of CRC in general.

Also, MS-MLPA technology was shown to be useful for testing CIMP. This will be essential if this molecular classification is to become of use in the clinical setting. Commonly used methylation techniques for determining CIMP are time consuming and they cannot be multiplexed. Moreover, they are all bisulphite treatment dependent, which in addition to requiring large amounts of DNA, introduces more variability. All the markers required for CIMP testing in CRC were multiplexed in a single MS-MLPA test that requires a small amount of DNA (as low as 50ng) without bisulphite treatment.

Finally, two of the investigated samples, CA248 and CA208 demonstrate the opposite ends of the CRC molecular background. CA248 did not have any defects in the investigated molecular characteristics, the patient was diagnosed at the age of 62, and had a very good survival. On the other hand CA208 had 5 somatic mutations in *APC*, *TP53* (2 mutations), *KRAS* and *BRAF*, in addition to, CIMP-L and CIN. The patient was diagnosed with CRC at the age of 51 and died within 1 month due to cancer progression. These 2 extreme cases demonstrate the prognostic value of molecular characterisation.

5. Focal Minimal Common Regions in Microsatellite Stable Colorectal Cancer

5.1 Introduction

Three main genomic instabilities are thought to drive CRC development; CIN, MSI and CIMP (Migliore *et al.* 2011). CIN is the most common and is associated with 65-85% of CRC cases (Derks *et al.* 2008, Issa 2008, Pino and Chung 2010, Migliore *et al.* 2011). Tumours that develop through the CIN pathway are characterised by frequent numerical and/or structural gains and losses of chromosomal segments or whole chromosomes (Rajagopalan *et al.* 2003). CIN is thought to drive CRC development through the amplification of oncogenes and the deletion of TSG (Kozma *et al.* 1994, Ozakyol *et al.* 2006, Tanaka *et al.* 2006, Martin *et al.* 2007, Ogino and Goel 2008, Migliore *et al.* 2011). This view is supported by the strong association reported between copy number abnormalities of cancer-related genes and their expression levels in CRC samples (Tsafrir *et al.* 2006, Sheffer *et al.* 2009, Brosens *et al.* 2010).

Although most of the chromosomal aberrations arise in a random fashion, some are recurrent and are commonly found in different colorectal tumours and other types of cancer (Martin *et al.* 2007, Beroukhim *et al.* 2010, Brosens *et al.* 2010). These common aberrations were probably selected for during tumour development for providing a survival advantage. Recurrent CNA provide the tumour with a way of targeting TSG and oncogenes to acquire one or more of the cancer hallmarks and drive tumourigenesis (Martin *et al.* 2007). Several common chromosomal abnormalities have been identified through conventional cytogenetic techniques, such as metaphase CGH and fluorescent in situ hybridisation (FISH), and are thought to drive CRC development (Hermsen *et al.* 2002, Diep *et al.* 2006, Lassmann *et al.* 2007). These chromosomal defects include gains of 8q, 13q and 20q and losses of 18q, 5q, 8p, 17q (Hermsen *et al.* 2002, Lassmann *et al.* 2007). However, due to their large size, identification of specific driver genes within these regions is usually challenging (Carvalho *et al.* 2009, Brosens *et al.* 2010).

Recent developments in the technologies used to define CIN, primarily array based methods, and advances in the computational algorithms used to interpret the results, have resulted in the acquisition of genome-wide information with high resolution (down to a few Kb) (Beroukhim *et al.* 2010, Brosens *et al.* 2010). This, in turn, has led to the identification of common focal chromosomal CNA. These focal CNA are usually smaller than 3Mb in size and thus contain a relatively small number of genes, hence simplifying the identification of driver genes (Brosens *et al.* 2010). Recently, focal CNA have led to the identification of several novel cancer driver genes with potential therapeutic and prognostic value in several cancer types including CRC (Andersen *et al.* 2010, Beroukhim *et al.* 2010, Bredel *et al.* 2010, Brosens *et al.* 2010, Poulogiannis *et al.* 2010a, Poulogiannis *et al.* 2010b, Veeriah *et al.* 2010, Dyrso *et al.* 2011). Nonetheless, the picture is far from complete and many of these recurrent CNA and their target genes are yet to be identified (Andersen *et al.* 2010).

Around 90.1% (n=48) of the 53 sporadic CRC cases investigated in this study were characterised by CIN and were MSS (Section 4.2.6.5). Only six of these CIN samples (~16%) were found to be CIMP-H. Therefore, CIN was considered the main form of genomic instability and a major driving force for CRC development amongst this cohort. The aims of the work described in this chapter were as follows. Firstly, identify the common CNA in the 45 CIN samples. Secondly, define the focal and highly recurrent minimal common regions (MCR) of the CNA and thirdly, identify any novel putative CRC driver genes located within these regions. A final aim was to confirm the aberrations in some of the candidate driver genes using qPCR.

5.2 Results

5.2.1 Focal amplifications and deletions

The aCGH data described in section 4.2.6 was used in the analysis performed in the following sections. Out of the 3097 CNA identified in the 45 chromosomally unstable samples, 1689 aberrations were focal (<3.0Mb) ranging in size from 0.0136Mb up to 2.9992Mb (mean: 0.9518Mb, median: 0.7058Mb). The focal aberrations consisted of 746 amplifications with a size range of 0.0136Mb-2.9897Mb (mean: 1.0882Mb, median: 0.9148Mb) and 943 deletions, with a size range of 0.02450Mb-2.9992Mb (mean: 0.8438Mb, median: 0.5747Mb).

5.2.2 Minimal common regions

The presence of several CNA arising at the same region in different samples allows definition of so-called “minimal common regions” (MCR). Figure 5.1 represents the concept of delimiting MCR by the combined analysis of several samples. The first step towards defining these MCR in the analysed cohort was attempted by using the context corrected common aberration (COCA) algorithm.

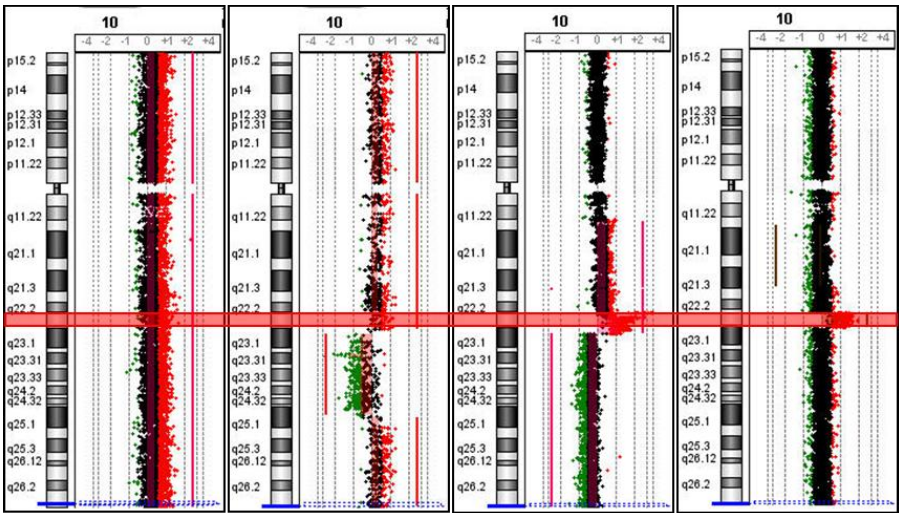


Figure 5.1 Delimiting an MCR from four different samples

Chromosome 10 aCGH data from 4 different cases, probes coloured in black represent genomic regions with a $\log_2$ratio around zero, probes coloured in red and green represent amplifications ($\log_2$ratio ≥ 0.5) and deletions ($\log_2$ratio ≤ -0.5) respectively. The 4 samples had amplifications of variable sizes. Combined analysis of the data delimits the minimal common region of amplification in the 4 cases. The MCR is marked by transparent red box.

5.2.2.1 Context corrected common aberration analysis results

In order to identify recurrent CNA, COCA analysis was performed as described in Section 2.2.15.11.4. COCA employs several steps. Firstly, the software considers all the aberrations reported by the ADM2 as candidate regions. Secondly, it assigns a score for each of these candidate regions. This score is based on the ADM2 results and the amplitude of the aberration, thus, it reflects the potential significance of each aberration in each sample. Finally, the hypothesis that a specific aberration is a common aberration in a certain number of samples is tested and each common aberration assigned a p-value and a final COCA score (-log10 of the p-value), which reflects the significance of the recurrent CNA.

When applied to the ADM2 results from the 45 cases, COCA identified 357 recurrent deletions and 286 recurrent amplifications. A graphical summary of the common CNA is presented in Figure 5.2.

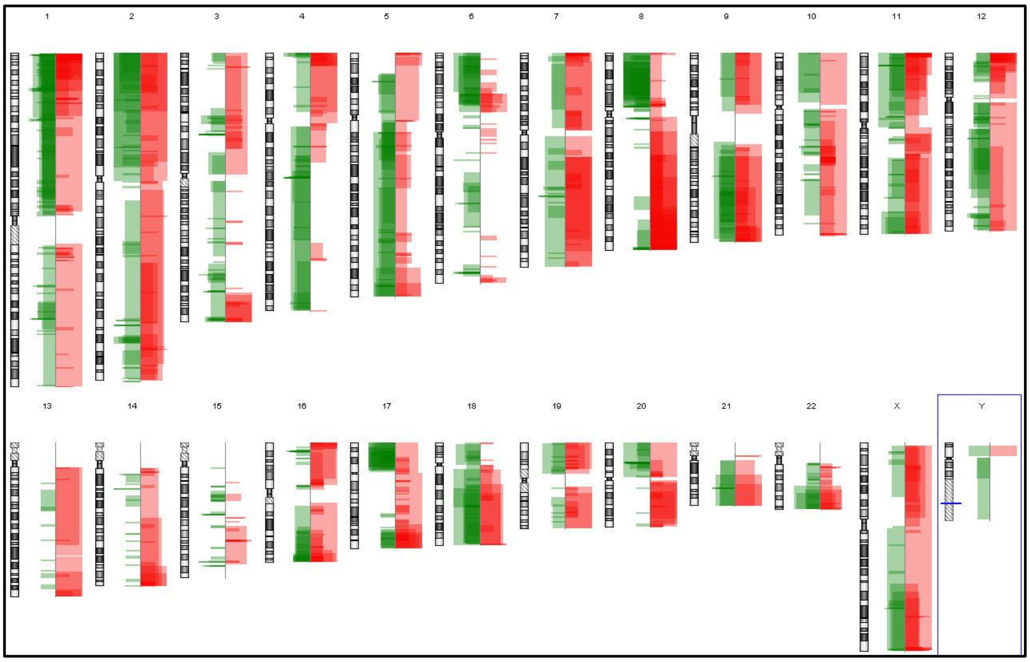


Figure 5.2 Heat map of common aberrations identified by COCA

Red represents amplifications, green represents deletions, the intensity of the colour reflects the frequency and the width of the bars reflects COCA score.

5.2.3 Focal minimal common regions

Some of the small common CNA occur within larger aberrant regions. In some of the cases, the COCA report summarises the large common aberrations encompassing the smaller ones, in other words, maximum rather than minimum common regions. Moreover, based on the amplitude and the size of the aberration from different samples, COCA separates a single common aberrant region into several (Figure 5.3 A). Therefore, the data output from COCA needed further analysis in order to accurately define the MCR.

The MCR were defined by applying an algorithmic approach that scanned the common regions identified by COCA output and recalculated the predicted MCR when necessary. The algorithm was designed by Wei-Yu Lin (Post-Doctoral researcher in our group) and delimits the MCR using aberration breakpoints in that region. The genomic location at the start of the aberration lying furthest downstream (the maximum “start” of the common aberrations in the area) was subtracted from the genomic location at the most upstream end of another aberration (the minimum “end” of the common aberrations in the area) (Figure 5.3 B).

All of the MCR identified by this algorithm were checked manually and aberrations that played a role in defining the size of the MCR, referred to as size determining events (SDE), were summarised (Figure 5.3B). Next, the MCR were filtered by size and only those with a size of 3Mbs or less were retained. These focal MCR (FMCR) were then filtered by frequency and any FMCR occurring in less than 10% of the cases were excluded from further analysis. The final step was to filter according to SDE, and FMCR were only kept when defined at least by 2 SDE (Figure 5.4). In order to increase the stringency of defining the significant FMCR, the minimum accepted COCA score was increased to ~2.0 (p-value of 0.01). These criteria resulted in a list containing 64 deletions and 32 amplifications (Table 5.1 and Table 5.2).

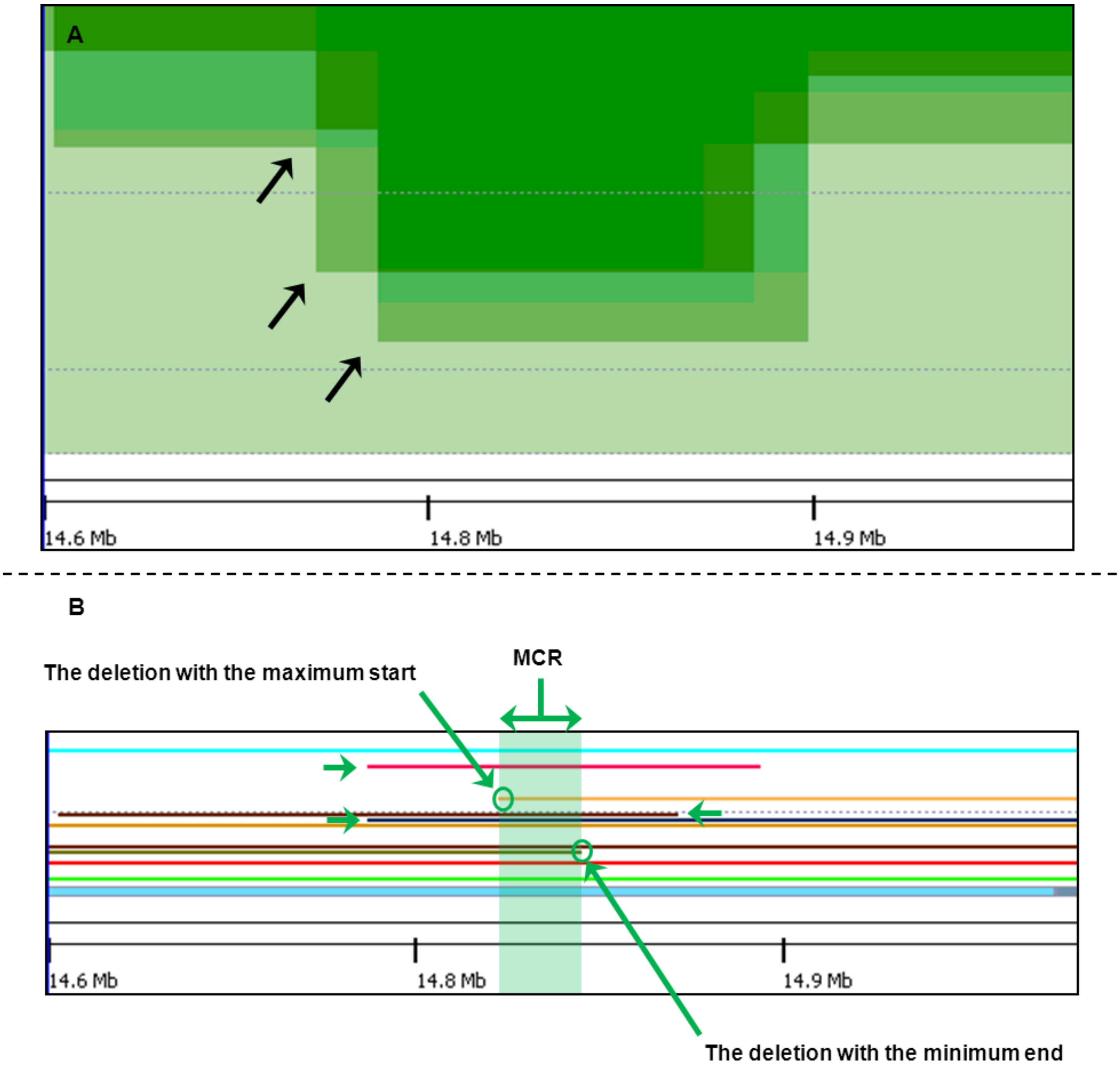


Figure 5.3 Recurrent CNA area on chr20p (*MACROD2* gene region)

A) COCA output identifies several common aberrations in the presented area. B) The coloured lines represent different deletions in the area. The green transparent box represents the MCR identified by the in-house algorithm. Green arrows mark all the SDE in that area. The green circles mark the genomic locations of the maximum start and the minimum end of the 2 SDE that defined the MCR size.

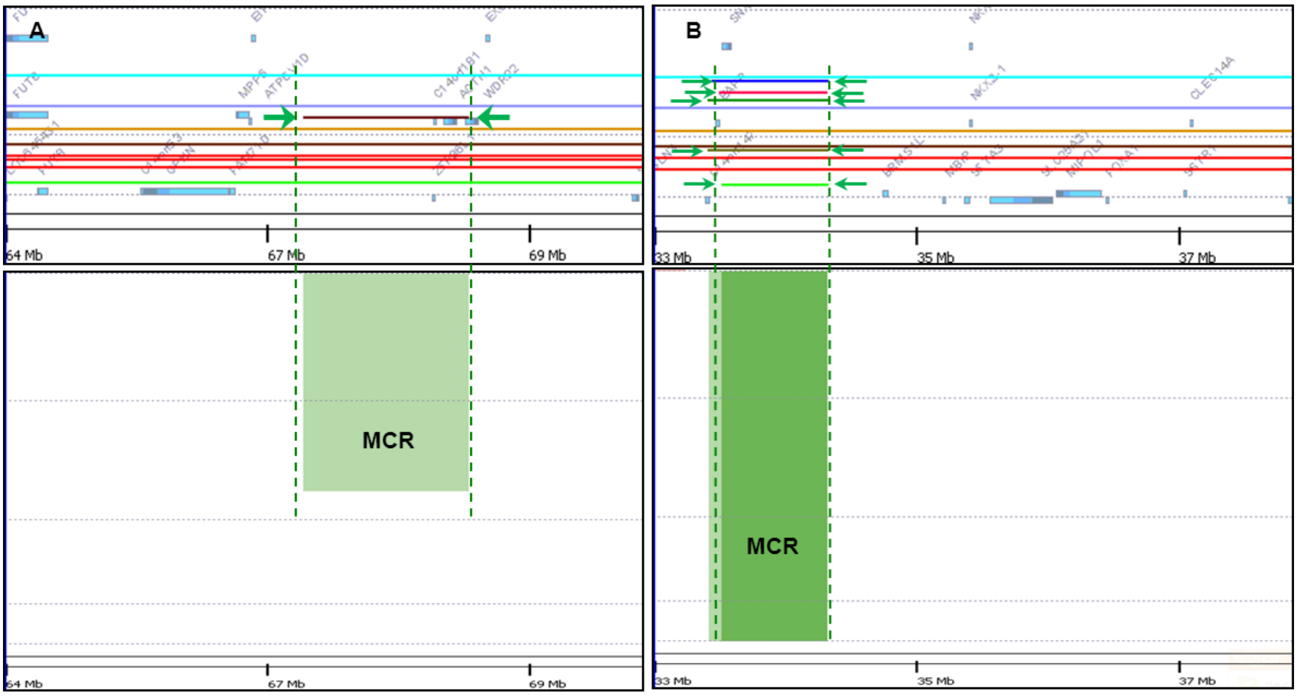


Figure 5.4 Size determining events

Coloured horizontal lines represent aberrations. SDE are marked by horizontal green arrows and vertical transparent green boxes represent MCR. A) An MCR identified by 1 SDE, B) An MCR identified by 5 SDE.

5.2.3.1 Summary of focal minimal common regions

The 64 deleted FMCR ranged in size between 0.03-2.64Mb (median: 0.42Mb, mean: 0.62Mb) and contained a total of 714 known genes with a range of 0-69 genes deleted per region (median: 4 genes, mean: 11 genes) (Table 5.1). The 32 amplified FMCR ranged between 0.03-1.96Mb in size (median: 0.83Mb, mean: 0.85Mb) and contained a total of 288 known genes with a range of 0-34 genes amplified per region (median: 3 genes, mean: 9 genes) (Table 5.2). It should be emphasised that for amplified FMCR, genes were only included in this count if they were fully contained within an FMCR. However, in the case of deleted FMCR, partially affected genes overlapping the size of delineation were included. Subsequently, deleted and amplified FMCR sets were checked for well-known cancer genes. To this end, the genes within the deleted and amplified FMCR were cross-checked against the complete gene list from the cancer gene census project (http://www.sanger.ac.uk/genetics/CGP/Census/, accessed August, 2011), and the CRC and breast cancer driver genes identified in a relatively recent high throughput re-sequencing study (Wood *et al.* 2007). This analysis indicated the presence of 25 "cancer genes" located within the deleted FMCR (~3.5% of the total number of deleted genes) and 11 "cancer genes" within the amplified FMCR (~3.8% of the total number of amplified genes).

The occurrence of a cancer gene within an FMCR does not necessarily imply that it is a driver gene within that area. For some "cancer genes", the type of FMCR (deleted or amplified) was not consistent with the known gene function, an example being the deletion of the known oncogene *NRAS* (Table 5.1). However, the gene function and the type of the FMCR were often consistent with expectation, examples include deletions of *MAP2K4* and *CDKN2C* (Table 5.1) and amplification of *FGFR1* (Table 5.2). Perhaps, most importantly, the classical *SMAD4* tumour suppressor deletion and oncogenic *MYC* amplification were both observed within deleted and amplified FMCR respectively. The frequency of both *SMAD4* deletions and *MYC* amplifications in the 45 chromosomally unstable cases was high at ~53% (Table 5.1 and Table 5.2).

Table 5.1 Deleted FMCR

Chromosomal location	Start	End	Size (Mb)	Recurrence (%)	SDE	Genes	Cancer genes
1p36.33	1479210	1851002	0.37	20.00	3	12	*MLLT11, ARNT*
1p36.11-p35.3	26462948	28551543	2.09	24.44	4	46	*ARID1A*
1p34.3	36375018	36734648	0.36	22.22	3	10	*THRAP3*
1p33	49486071	50272200	0.79	22.22	3	1	
1p33-p32.3	50698666	51354770	0.66	20.00	2	2	*CDKN2C*
1p31.1	78062885	78218246	0.16	20.00	2	3	
1p21.1	102307656	103278518	0.97	20.00	2	1	*COL11A1*
1p13.2	112844021	115479162	2.64	17.78	2	29	*NRAS, TRIM33*
1q21.2-q21.3	147644631	149711565	2.07	11.11	3	69	
1q31.3	196799738	196948778	0.15	13.33	2	1	
2p14	69914199	70385803	0.47	11.11	4	10	
2q33.1	197546654	198133273	0.59	13.33	4	6	
2q33.1	201432284	201850246	0.42	17.78	6	9	
2q37.1	232634311	232902204	0.27	11.11	2	1	
3p14.3	57521404	57652691	0.13	24.44	4	3	
3p14.2	60078018	61195823	1.12	31.11	7	1	*FHIT*

3q21.3	129834187	130497580	0.66	13.33	2	11	*RPN1, CNBP*
3q26.1	163997028	164101976	0.10	17.78	2	0	
3q26.31	176177125	176426607	0.25	13.33	3	1	
4q22.1	91340468	92674544	1.33	33.33	8	2	
4q35.1	185447304	186062400	0.62	31.11	3	5	
5q12.1	59145099	59183979	0.04	22.22	2	1	
5q13.2	68438765	68906238	0.47	24.44	2	14	
5q35.1-q35.2	172069779	172570325	0.50	17.78	3	7	
6p22.1	26141996	26381758	0.24	15.56	2	26	
6p21.33-p21.32	31388886	32258223	0.87	13.33	3	63	
6p21.31	35536901	35954862	0.42	11.11	2	11	*FANCE*
6q21	111214198	111443834	0.23	15.56	2	4	
6q26	162357125	163049854	0.69	22.22	7	1	
7q22.1	99307476	101900781	2.59	17.78	5	68	*CUX1*
7q22.2	105680893	105712574	0.03	11.11	3	1	
8p22	16198781	16863085	0.66	48.89	2	0	
8p11.23	39431213	39480174	0.05	22.22	3	0	
8q11.21	48247963	48556480	0.31	13.33	4	1	
8q24.3	145103393	145715756	0.61	13.33	3	36	*RECQL4*
9q21.13	73537301	74412105	0.87	17.78	3	7	
9q22.2	91117607	91411397	0.29	22.22	2	5	
10q23.32-q23.33	94363548	94454268	0.09	22.22	3	2	
11p15.4	3712771	3873442	0.16	15.56	3	4	*NUP98*
11p15.4	9172449	9363610	0.19	22.22	5	3	
11p11.2	47337056	47964454	0.63	11.11	4	14	
11p11.2-p11.12	48215588	50164184	1.95	13.33	5	9	
11q13.1	63705254	63872837	0.17	15.56	2	17	
12p13.31	6314214	6963027	0.65	11.11	4	37	*ZNF384, ATN1*
12q14.2	63277389	63368109	0.09	20.00	4	1	
12q24.31	122199734	122593396	0.39	17.78	5	8	
14q12	23615376	23888497	0.27	24.44	3	28	
14q13.2	34108596	34950339	0.84	28.89	5	9	
15q21.3	53286334	53506655	0.22	28.89	2	4	
15q26.1	88098779	89290699	1.19	31.11	4	20	
15q26.1	91163633	91270553	0.11	28.89	3	1	*BLM, CRTC3, IDH2*
16p13.3-p13.2	6132536	7018275	0.89	24.44	7	1	
17p13.1	8743396	8866516	0.12	46.67	2	1	
17p13.1-p12	10957541	12400968	1.44	48.89	5	3	*MAP2K4*
18p11.21	11759039	12046289	0.29	42.22	2	4	
18q12.2	32736160	32823241	0.09	51.11	2	1	
18q21.2	46648889	46859053	0.21	53.33	3	3	*SMAD4*
19p13.3	909095	1223851	0.31	20.00	4	16	*STK11*
19q13.11	17936643	18549005	0.61	20.00	2	21	*ELL*
20p12.1	14376202	16071135	1.69	37.78	10	1	
21q22.11	33554005	33651205	0.10	28.89	7	3	
22q13.1	37647505	38597446	0.95	26.67	3	18	*PDGFB*
22q13.33	48580338	49052472	0.47	26.67	3	16	
Xq25	122938369	123237260	0.30	15.56	3	1	

Table 5.1 summarises the deleted FMCR identified in the 45 chromosomally unstable samples. FMCR are arranged by genomic location (hg18 assembly). Genes within the FMCR were determined according to the UCSC genome browser (http://genome.ucsc.edu/, accessed August, 2011). “Cancer genes” are derived from the cancer gene census and Wood *et al.* 2007.

Table 5.2 Amplified FMCR

Chromosomal location	Start	End	Size (Mb)	Recurrence (%)	SDE	Genes	Cancer genes
1p36.32	2412144	3630036	1.22	26.67	7	11	*PRDM16, TNFRSF14*
1p36.13	17783440	18947883	1.16	17.78	4	4	*PAX7*
1p35.1-p34.3	33866043	34889487	1.02	15.56	4	2	
1p34.3	36881028	37535891	0.65	20.00	5	1	
1q21.3	150753997	150969016	0.22	22.22	5	19	
1q21.3-q22	153171645	153567906	0.40	20.00	4	29	*MUC1, THBS3*
1q32.1	203312447	205274691	1.96	15.56	2	34	*ELK4, SLC45A3*
2p13.2-13.1	72867953	73995941	1.13	17.78	3	16	
2q11.2	96118572	97026741	0.91	15.56	3	21	*CNNM4*
2q14.2	120835644	121713081	0.88	15.56	3	1	
2q21.2	132747027	132833787	0.09	17.78	2	0	
2q33.1	199239915	200062504	0.82	15.56	2	1	
2q35	218354227	218556081	0.20	20.00	5	1	
3p25.3	10406898	10978905	0.57	15.56	2	1	
5p15.33	131946	1578616	1.45	26.67	2	22	*SLC6A3*
6p21.2-p21.1	39059270	40756937	1.70	24.44	2	9	
7p22.3-p22.2	1262920	2812963	1.55	46.67	2	18	
7p22.1-p21.3	7033162	7829887	0.80	48.89	5	4	
8p12-p11.23	38303146	39234126	0.93	15.56	5	7	*FGFR1*
8q23.3	113827791	114547454	0.72	51.11	4	0	
8q24.21	128331769	128940144	0.61	53.33	2	1	*MYC*
10q22.3	78238542	79084656	0.85	20.00	4	1	
10q26.3	133770546	135084977	1.31	15.56	3	22	
11p15.5-p15.4	916837	2837148	1.92	33.33	10	31	
12p13.32-p13.31	4282429	5364999	1.08	40.00	4	12	
14q11.2	21680489	21935641	0.26	20.00	3	0	
15q21.3-q22.1	55517770	56143371	0.63	11.11	2	2	
15q24.1	72478743	72569181	0.09	17.78	2	2	
16p13.3	738567	1229020	0.49	28.89	4	13	
17q11.2-q12	28205256	29488626	1.28	28.89	3	2	
22q11.21	18517833	18686313	0.17	20.00	4	1	
22q13.31	44865639	44898001	0.03	17.78	3	0	

Table 5.2 summarises the amplified FMCR identified in the 45 chromosomally unstable samples. FMCR are arranged by genomic location (hg18 assembly). Genes within the FMCR were determined according to the UCSC genome browser (http://genome.ucsc.edu/, accessed August, 2011). “Cancer genes” are derived from cancer gene census and Wood *et al.* 2007.

5.2.3.2 FMCR and published focal aberrations in CRC

In order to further validate our panel of deleted and amplified FMCR, they were compared to previously identified focal aberrations in CRC samples. Common chromosomal aberrations identified in four recent studies on CRC were used for the comparison (Martin *et al.* 2007, Leary *et al.* 2008, Andersen *et al.* 2010, Brosens *et al.* 2010). A summary of the technologies used in identifying the chromosomal aberrations, the platforms and the resolutions, in addition to the number of primary CRC cases and CRC cell lines included in each study is presented in Table 5.3.

Table 5.3 Previously published studies on focal aberrations in CRC

	Martin *et al.* 2007	Leary *et al.* 2008	Brosens *et al.* 2010	Andersen *et al.* 2010
Technology	aCGH	SNP array	aCGH	SNP array
Platform	Agilent 1AV2	SNP6.0	Agilent 44K	SNP6.0
Resolution	~55Kb	~30-50Kb	~43Kb	~30-50kb
CRC Cases	42	36	38	33
CRC cell lines	37	NA	NA	40
Adenomas	NA	NA	NA	40

In order to perform a valid comparison, only focal aberrations ≤3Mb in size were included. The combined total of focal aberrations identified in these studies was 187 amplifications and 189 deletions. The published focal aberrations from each study were compared to identify any levels of similarity. The 4 studies had 10 overlapping focal deletions (5.3%) and 29 overlapping focal amplifications (15.5%). The published aberrations were then compared to the FMCR identified in our study. In total, 27 overlapped focal deletions (42.2% of our identified deleted FMCR) and 11 overlapped focal amplifications (34.4% of our identified amplified FMCR) were observed.

The common overlapping areas between our FMCR and the published focal aberrations were investigated in order to check whether they were biologically significant. This was performed by examining if the overlapped regions contained the known cancer genes. Four of the overlapped FMCR carried established cancer genes (*MYC*, *FGFR1*, *SMAD4* and *MAP2K4*). The overlapped regions were smaller in all cases compared to identified FMCR, and they all contained these cancer genes. Interesting examples were *FGFR1* regions, where an overlap between our identified amplification FMCR of 931Kb (containing 7 genes) (Table 5.2), and a published focal amplification of 1475Kb resulted in a delimited region of 145Kb that contained only 2 genes, one of which was *FGFR1*.

5.2.3.3 FMCR and published CNA across several cancer types

A large somatic CNA study was recently performed in 26 different cancer types, including CRC (Beroukhim *et al.* 2010). The study identified a list of the 20 most common somatic deletions and amplifications across the analysed cancer types. Moreover, a candidate driver gene was also selected for each of the common aberrations (Table 5.4).

A comparison between our FMCR and the most common regions in the above study revealed an overlap with five of the deletion areas (25%) and 3 of the amplification areas (15%). All the candidate genes identified by Beroukhim and colleagues in their study were contained within the overlapping areas from our FMCR.

Table 5.4 Candidate driver genes of the most common cancer CNA

Deletion		**Amplification**	
Chromosome	**Gene**	**Chromosome**	**Gene**
1p	*DFFB*	1p	*MYCL1*
2q	*LRP1B*	1q	*MCL1*
2q	*BOK*	3q	*PRKC1*
3p	***FHIT***	**5p**	***TERT***
4q	***TMSL3***	7p	*EGFR*
4q	*FAT1*	**8p**	***FGFR1***
5q	*PLK2*	8p	*IKBKB*
6q	***PARK2***	**8q**	***MYC***
7q	*TRB*	11q	*CCND1*
8p	*CSMD1*	12p	*KRAS*
9p	*PTPRD*	12q	*CDK4*
9p	*CDKN2A/B*	12q	*HMGA2*
10q	*PTEN*	12q	*MDM2*
12p	*ETV6*	12q	*CCT2*
13q	*RB1*	14q	*NKX2-1*
16p	***A2BP1***	15q	*IGF1R*
16q	*WWOX*	17q	*ERBB2*
19q	*PPAP2C*	19q	*CCNE1*
20p	***MACROD2***	22q	*CRKL*
22q	*PPM1F*	Xq	*IRAK1*

Table 5.4 summarises the candidate driver genes identified across 26 different cancer types. Genes in bold were also found in the FMCR identified in this study (Beroukhim *et al.* 2010).

5.2.3.4 FMCR and pathway analysis

In order to search for any specific patterns or pathways affected by the genes within the deleted or amplified FMCR, the Database for Annotation, Visualisation, and Integrated Discovery (DAVID) v6.7 was used (http://david.abcc.ncifcrf.gov/, accessed August, 2011). DAVID is a high throughput data mining software that can be used to analyse large gene lists (Huang da *et al.* 2009). DAVID performs enrichment analysis (based on biological annotations for the target genes) to identify any biological pathways that are statistically significantly over-represented in the analysed gene list (Huang da *et al.* 2009). The identified genes within the deleted and the amplified FMCR were analysed separately.

5.2.3.4.1 Deleted FMCR pathway analysis

For the deleted regions, the most enriched cancer-related pathway was Apoptosis (p-value = 0.014) (Figure 5.5). Ten apoptotic genes (*CASP3*, *CASP8*, *CASP10*, *CFLAR NFKBIA*, *BAD*, *PIK3R2*, *PIK3R5*, *TNFRSF1A* and *TNF*) were found to be within the deleted FMCR. Seven of these genes have pro-apoptotic roles and 3 are anti-apoptotic (Table 5.5).

Table 5.5 Deleted apoptotic genes

Proapoptotic		Antiapoptotic	
Gene	**Frequency% (n)***	**Gene**	**Frequency% (n)***
CASP3	31.1 (14)	*PIK3R5*	46.7 (21)
NFKBIA	28.9 (13)	*PIK3R2*	20.0 (9)
CASP8	17.8 (8)	*CFLAR*	17.8 (8)
CASP10	17.8 (8)		
BAD	15.6 (7)		
TNF	13.3 (6)		
TNFRSF1A	11.1 (5)		

* Frequency out of the 45 chromosomally unstable tumours

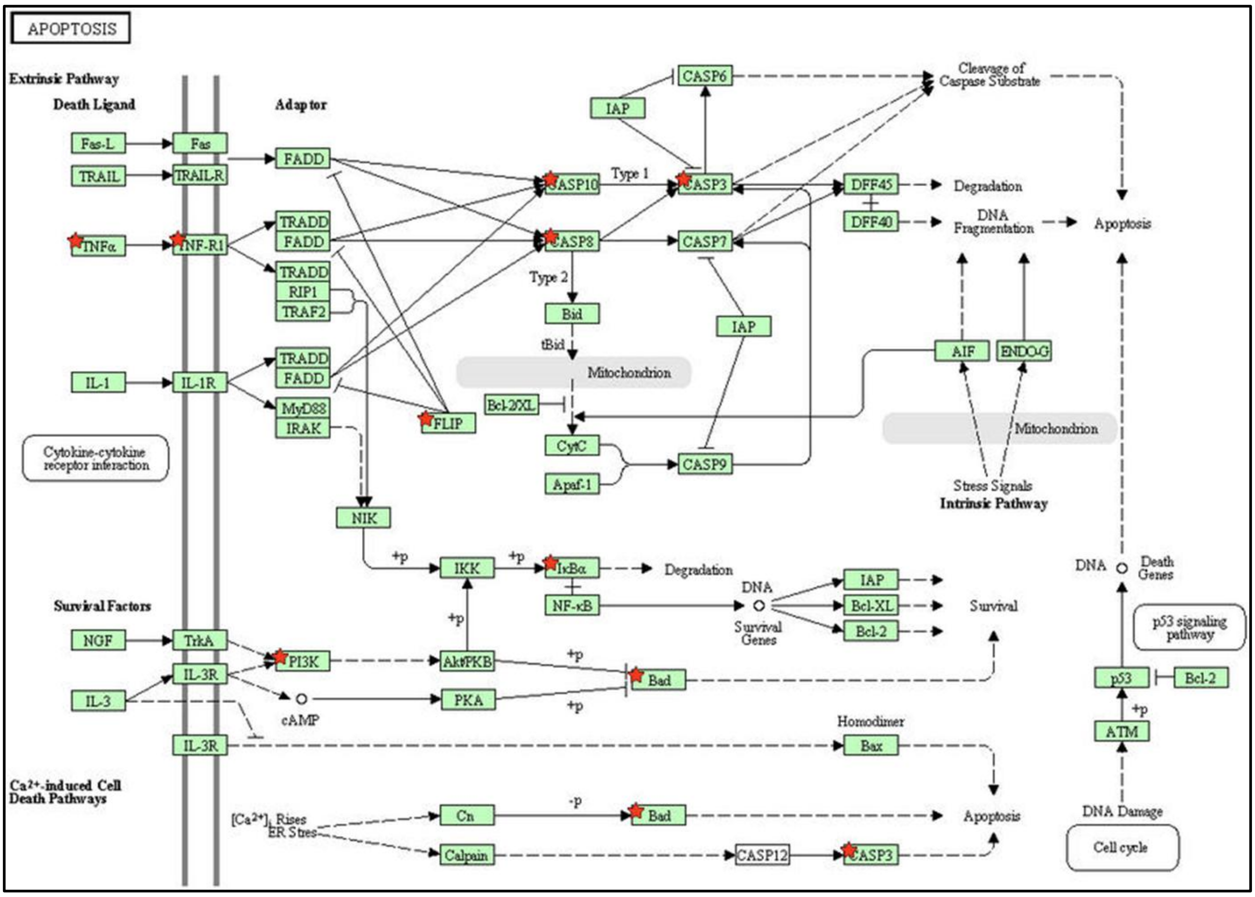


Figure 5.5 Apoptotic genes in the deleted FMCR (DAVID output)

DAVID output showing the apoptosis signalling pathway with deleted genes marked by a red star.

Another cancer related pathway that was found to be enriched within deleted FMCR was the P53 signalling pathway (p-value = 0.079). In addition to the apoptotic genes *CASP8* and *CASP10*, 5 more genes from this pathway were also deleted; *CCNB1*, *GADD45G*, *SERPINE1*, *SESN2* and *SFN*, all of which have reported anti-survival functions (Figure 5.6).

Table 5.6 Deleted P53 pathway genes

Gene	Frequency% (n)*
CCNB1	24.4 (11)
SESN2	24.4 (11)
SFN	24.4 (11)
GADD45G	20.0 (9)
SERPINE1	15.6 (7)

* Frequency out of the 45 chromosomally unstable tumours

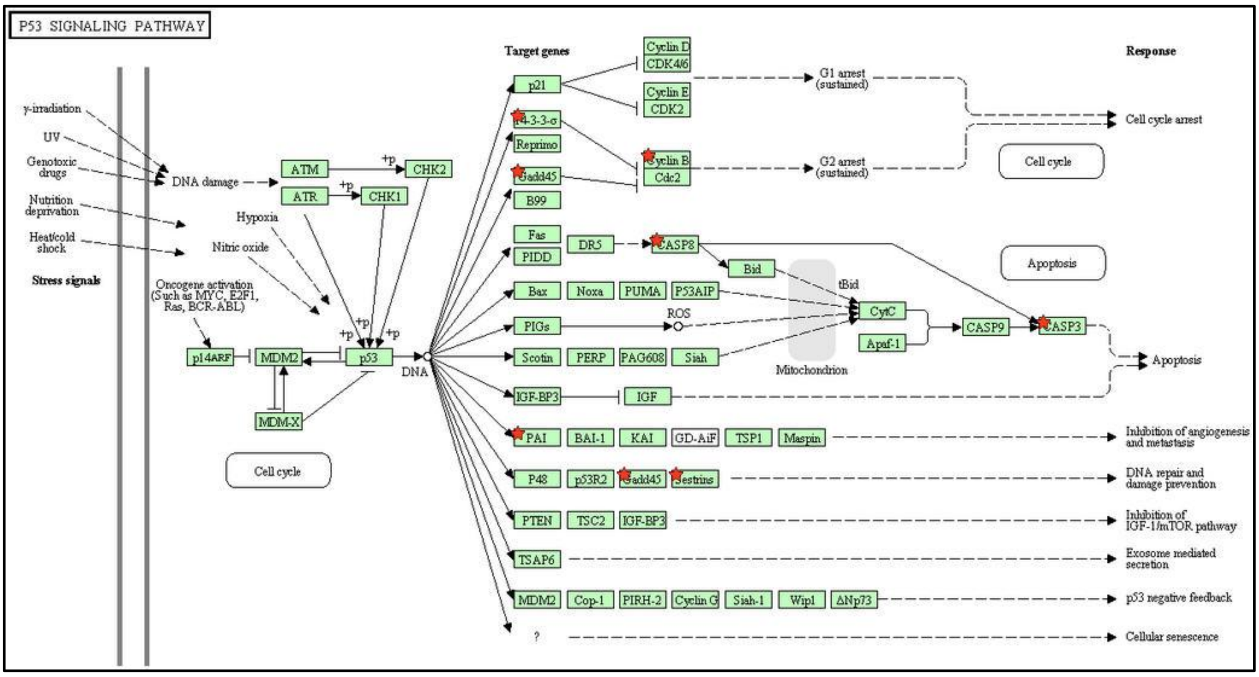


Figure 5.6 P53 signalling pathway and the deleted FMCR (DAVID output)

DAVID output showing the P53 signalling pathway, with deleted genes marked by a red star.

5.2.3.4.2 Amplified FMCR pathway analysis

For amplified FMCR, the oncogenic MAPK signalling pathway was the only statistically significantly over-represented pathway (p-value = 0.032) (Figure 5.7). Nine genes in the MAPK pathway (*CACNA1H*, *DUSP2*, *DUSP8*, *FGF23*, *FGF6*, *FGFR1*, *NTF3*, *ELK4*, *MAPKAPK2* and *MYC*) were found to be within the amplified FMCR. Seven of these genes are already known to promote growth and survival (Table 5.7).

Table 5.7 Amplified MAPK genes

Pro-survival		Anti-survival	
Gene	**Frequency % (n)**	**Gene**	**Frequency % (n)**
MYC	53.3 (24)	*DUSP8*	33.3 (15)
FGF23	40.0 (18)	*DUSP2*	15.6 (7)
FGF6	40.0 (18)		
MAPKAPK2	15.6 (7)		
ELK4	15.6 (7)		
CACNA1H	28.9 (13)		
FGFR1	15.6 (7)		

* Frequency out of the 45 chromosomally unstable tumours

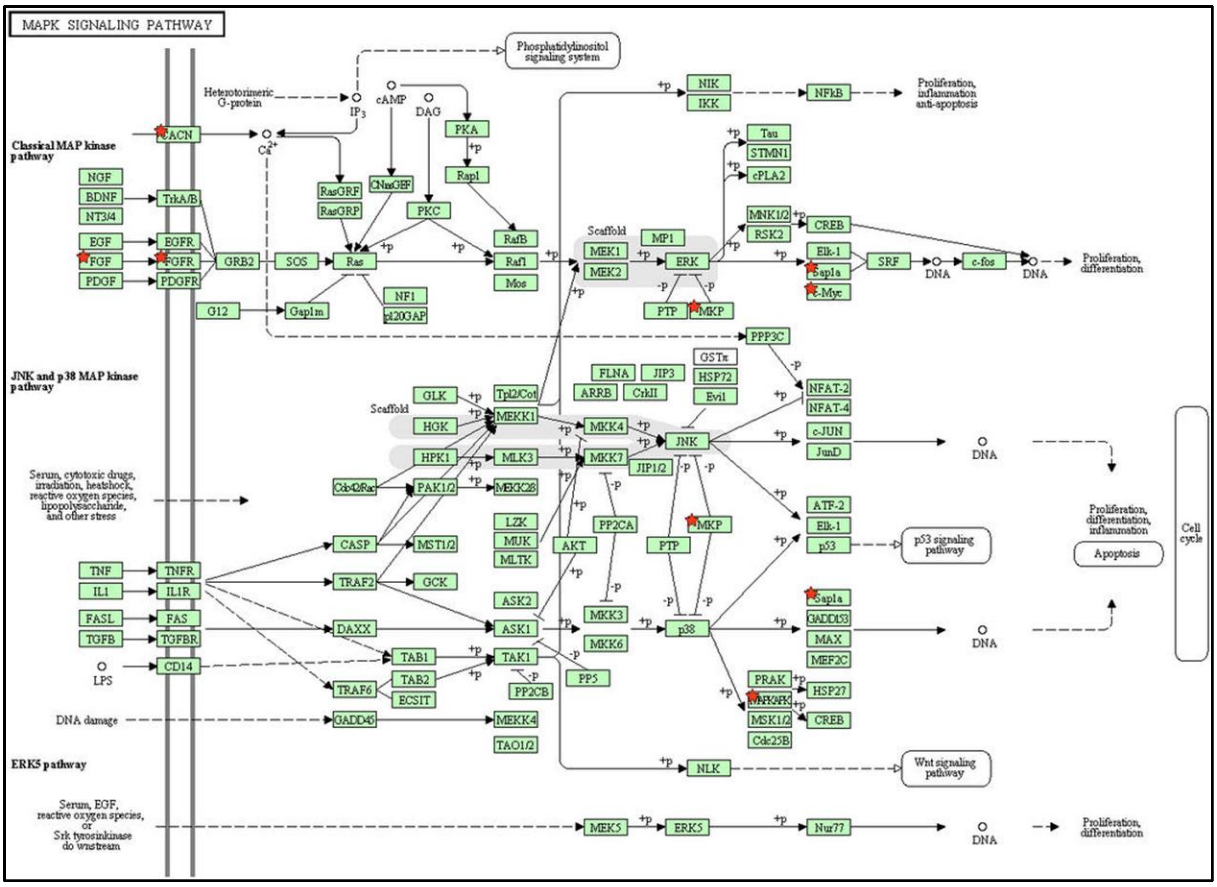


Figure 5.7 MAPK pathway defects

DAVID output showing the MAPK signalling pathway with amplified genes marked by a red star.

5.2.3.5 Candidate FMCR and genes

In order to reduce our list of FMCR and identify strong candidate driver genes for follow up studies, the FMCR criteria described in Section 5.2.3 were made more stringent. Firstly, candidate FMCR were defined by 4 SDE, instead of 2 (i.e. ~10% of the chromosomally unstable samples had targeted CNA in the region). Secondly, instead of defining "focal" by size only, a candidate FMCR was considered if it had a maximum of 12 genes in the aberrant area (i.e. ≤3.0Mb and ≤12 genes). This modification was introduced to help focus the search for candidate driver genes. The more stringent FMCR definition reduced the number of candidate FMCR to 17 deletions (Table 5.8) and 11 amplifications (Table 5.9). The deleted regions had a size range of 0.09-1.95Mb (mean: 0.74Mb, median: 0.59Mb) and the total number of genes in these areas was 71 with a range of 1-10 genes (median: 3, mean: 4.2) (Table 5.8). The amplified regions had a size range of 0.17-1.22Mb (mean: 0.80Mb, median: 0.85Mb) and a total of 44 genes with a range of 0-12 (mean: 4, median: 2) (Table 5.9).

Table 5.8 Shortlist of deleted FMCR

Chromosomal location	Start	End	Size (Mb)	Recurrence (%)	SDE	Genes	Candidate genes
2p14	69914199	70385803	0.47	11.11	4	10	*GMCL1, MAD1, PCBP1*
2q33.1	197546654	198133273	0.59	13.33	4	6	
2q33.1	201432284	201850246	0.42	17.78	6	9	*CASP8, CASP10*
3p14.3	57521404	57652691	0.13	24.44	4	3	
3p14.2	60078018	61195823	1.12	31.11	7	1	*FHIT*
4q22.1	91340468	92674544	1.33	33.33	8	2	*TMSL3*
6q26	162357125	163049854	0.69	22.22	7	1	*PARK2*
8q11.21	48247963	48556480	0.31	13.33	4	1	
11p15.4	9172449	9363610	0.19	22.22	5	3	
11p11.2-p11.12	48215588	50164184	1.95	13.33	5	9	*FOLH1*
12q14.2	63277389	63368109	0.09	20.00	4	1	
12q24.31	122199734	122593396	0.39	17.78	5	8	*CDK2AP1*
14q13.2	34108596	34950339	0.84	28.89	5	9	*NFKBIA*
16p13.3-p13.2	6132536	7018275	0.89	24.44	7	1	*A2PB1*
17p13.1-p12	10957541	12400968	1.44	48.89	5	3	*MAP2K4*
20p12.1	14376202	16071135	1.69	37.78	10	1	*MACROD2*
21q22.11	33554005	33651205	0.10	28.89	7	3	*IFNAR1, IFNAR2*

Table 5.9 Shortlist of amplified FMCR

Chromosomal location	Start	End	Size (Mb)	Recurrence (%)	SDE	Genes	Candidate genes
1p36.32	2412144	3630036	1.22	26.67	7	11	*PRDM16*
1p36.13	17783440	18947883	1.16	17.78	4	4	*PAX7*
1p35.1-p34.3	33866043	34889487	1.02	15.56	4	2	
1p34.3	36881028	37535891	0.65	20.00	5	1	
2q35	218354227	218556081	0.20	20.00	5	1	*TNS1*
7p22.1-p21.3	7033162	7829887	0.80	48.89	5	4	*RPA3*
8p12-p11.23	38303146	39234126	0.93	15.56	5	7	*FGFR1*
8q23.3	113827791	114547454	0.72	51.11	4	0	NA
10q22.3	78238542	79084656	0.85	20.00	4	1	*KCNMA1*
12p13.32-p13.31	4282429	5364999	1.08	40.00	4	12	*FGF23, FGF6*
22q11.21	18517833	18686313	0.17	20.00	4	1	

5.2.3.5.1 Deleted FMCR

In total 71 genes were affected by the 17 deleted FMCR. Sixteen of these genes (~22.5%) were cancer-related with known or potential anti-tumourigenesis activity (Table 5.8). Five of these genes (*FHIT*, *TMSL3*, *PARK2*, *A2BP1* and *MACROD2*) were within the 20 most common deleted CNA in CRC and various other cancer types (Section 5.2.3.3) (Beroukhim *et al.* 2010). *PARK2* is an interesting example as it was initially identified as a candidate TSG in glioblastoma multiforme (GBM) and

CRC through focal MCR (Veeriah *et al.* 2010) and *in vitro* and *in vivo* experiments have subsequently confirmed its anti-tumour activity in CRC (Poulogiannis *et al.* 2010b). *MAP2K4* was also reported to be commonly deleted in CRC (Leary *et al.* 2008). None of the other genes have previously been reported to be deleted in CRC. The following sections briefly describe the candidate genes identified in the deleted FMCR.

5.2.3.5.1.1 2p14

This region contained 3 candidate CRC TSG; *GMCL1*, *MAD1* and *PCBP1*. Germ cell-less homolog 1 (*GMCL1*) was previously proposed as a TSG as it was shown to increase the transcriptional activity of P53 (Masuhara *et al.* 2003). *MAD1* (encoding MAX dimerization protein 1), also known as *MXD1*, was shown to have anti-tumour activity and its protein was shown to be under-expressed in several human cancers including CRC (Gunes *et al.* 2000, Toaldo *et al.* 2010). Finally, PolyC-RNA-binding protein 1 (PCBP1) was shown to inhibit the oncogenic kinase AKT in CRC (Wang *et al.* 2010).

5.2.3.5.1.2 2q33.1

Deletion of this region has been previously implicated in several cancers, including gastric and lung cancer and neuroblastoma (Otsuka *et al.* 1996, Nishizuka *et al.* 1998, Teitz *et al.* 2000, Takita *et al.* 2001, Shivapurkar *et al.* 2002a, Geelen *et al.* 2004). The region contains the well characterised pro-apoptotic genes *CASP8* and *CASP10* (Zhang *et al.* 2004). *CASP8* inactivating mutations were previously reported in ~5% of advanced CRC cases (Kim *et al.* 2003).

5.2.3.5.1.3 11p11.2-p11.12

Folate hydrolase gene (*FOLH*1) was considered the candidate driver gene in this deleted FMCR. *FOLH1* plays a role in folic acid absorption by hydrolysing folate bound to polygultamate. Lower levels of folic acid are thought to increase the risk of

CRC and to affect the efficiency of some chemotherapeutic agents (Giovannucci *et al.* 1995, DeVos *et al.* 2008, Sadahiro *et al.* 2010, Porcelli *et al.* 2011).

5.2.3.5.1.4 12q24.31

This region contains a candidate TSG which encodes the cell cycle regulator protein cdk2-associating protein1 (CDK2AP1). *CDK2AP1* was recently shown to have anti-metastatic activities in squamous cell carcinoma and prostate cancer (Zolochevska and Figueiredo 2010a, Zolochevska and Figueiredo 2010b).

5.2.3.5.2 14q13.2

NFKBIA (encoding nuclear factor of κ-light polypeptide gene enhancer in B-cells inhibitor-α) is the candidate TSG in this region. *NFKBIA* was recently reported to be deleted in GBM (Bredel *et al.* 2010). Over-expression of *NFKBIA* in GBM cell lines had anti-growth, anti-survival and anti-metastatic activities (Bredel *et al.* 2010). Moreover, inhibiting *NFKBIA* expression in lung cancer cell lines was shown to induce resistance to chemotherapeutic agents (Bivona *et al.* 2011). Interestingly, *NFKBIA* SNPs were also previously associated with CRC risk (Gao *et al.* 2007).

5.2.3.5.2.1 21q22.11

Interferon receptors genes (*IFNAR1 and 2*) were predicted to be TSG because their protein levels were shown to be downregulated in bladder cancer, and correlated with advanced stages and resistance to chemotherapy (Zhang *et al.* 2010).

5.2.3.5.3 Amplified FMCR

In total, 44 genes were within the amplified FMCR. Eight (18.2%) of these genes were known or predicted to have oncogenic activities (Table 5.9). *FGFR1* was within the 20 most common amplified CNA in CRC and various other cancer types (Section 5.2.3.3). *FGF6* and *FGF23* were previously reported to be amplified in CRC (Sheffer *et al.* 2009). The rest of the genes were not previously reported to be amplified in CRC. The following sections briefly describe the candidate genes in the amplified FMCR.

5.2.3.5.3.1 1p36.32

PR domain-containing 16 gene (PRDM16) is a known oncogene implicated in acute myeloid leukaemia (AML) and osteosarcoma (Man *et al.* 2004, Shing *et al.* 2007)

5.2.3.5.3.2 1p36.13

This region contains the transcription factor paired box gene 7 (*PAX7*). *PAX7* is known to play an important role in alveolar rhabdomyosarcoma by forming a fusion protein with forkhead in rhabdomyosarcoma gene (*FKHR*) (Barr *et al.* 1996). *PAX7* is amplified in 20% of the alveolar rhabdomyosarcoma cases with the fusion gene (Barr *et al.* 1996).

5.2.3.5.3.3 2q35

Tensin 1 gene (*TNS1*) is the only gene in this region. *TNS1* overexpression *in vitro* was previously shown to significantly promote cell migration in fibroblasts (Chen *et al.* 2002).

5.2.3.5.3.4 7p22.1-p21.3

Replication protein A3 gene (*RPA3*) located in this area was recently shown to be amplified in metastatic melanoma (focal MCR) and to play an essential role in tumour invasion (Kabbarah *et al.* 2010).

5.2.3.5.3.5 10q22.3

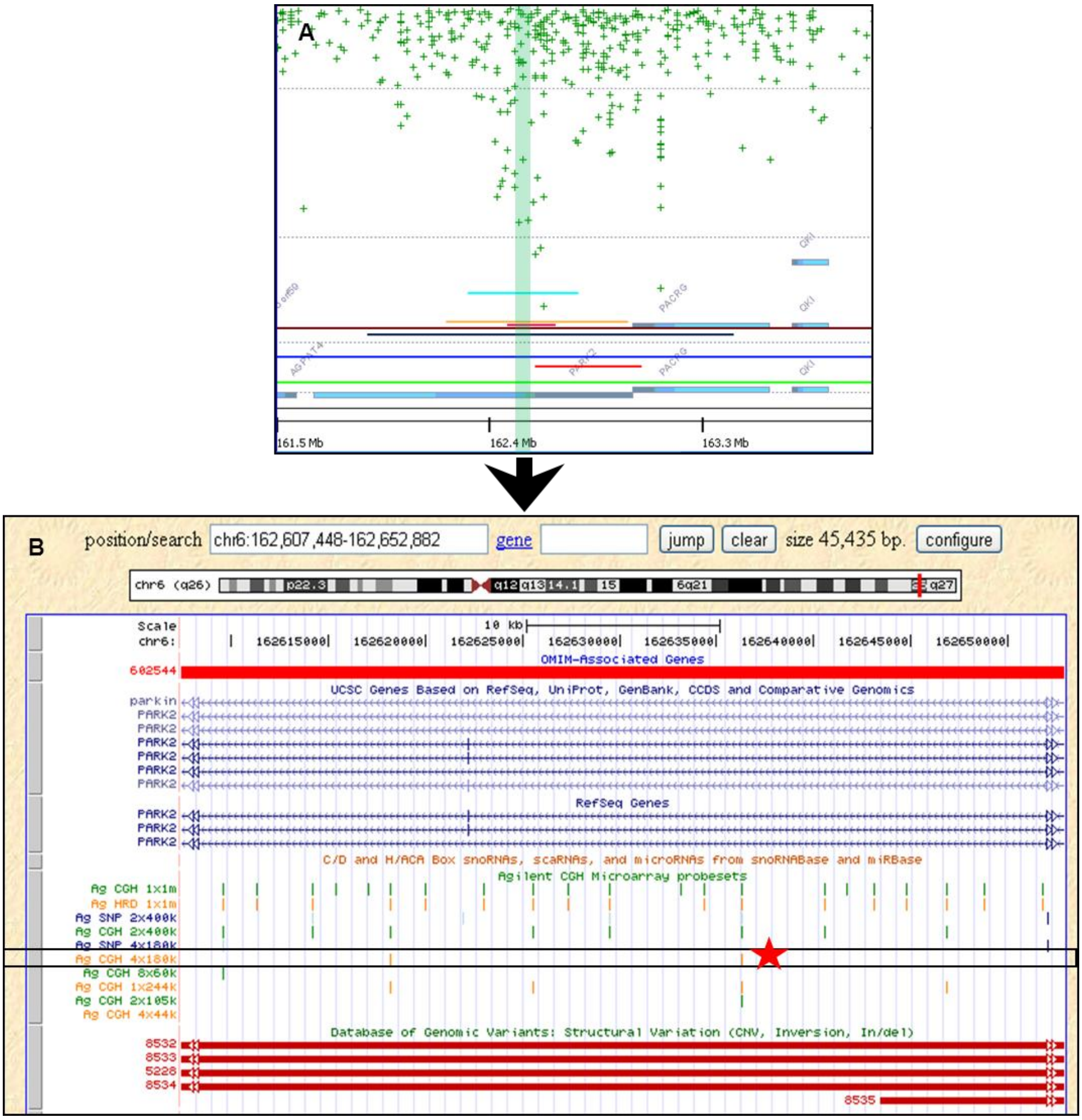
Large conductance calcium-activated potassium channel alpha subunit gene (*KCNMA1*), the only gene in this area, was previously shown to be amplified in prostate cancer cases and overexpressed in metastatic breast cancer (Bloch *et al.* 2007, Khaitan *et al.* 2009). Functional *KCNMA1* assays confirmed a role in cancer cell growth and invasion *in vitro* (Bloch *et al.* 2007, Khaitan *et al.* 2009).

5.2.3.6 FMCR confirmation using TaqMan copy number assays

The reproducibility of the aCGH experiments (Section 4.2.6.4) was confirmed by performing duplicate experiments. In order to check the accuracy of the aCGH platforms in detecting these focal aberrations, TaqMan copy number assays (qPCR) were used (Section 2.2.9.2). Three candidate genes (*PARK2*, *NFKBIA* and *KCNMA1*) that occur within the shortlisted FMCR, were chosen to be verified by qPCR. *PARK2* and *NFKBIA* reside within deleted FMCR (6q26 and 14q13.2 respectively) (Table 5.8) and *KCNMA1* resides within an amplified FMCR (10q22.3) (Table 5.9). *PARK2* was chosen because deletions in this gene were recently observed (and also confirmed) to occur in CRC and other cancers (Beroukhim *et al.* 2010, Poulogiannis *et al.* 2010b, Veeriah *et al.* 2010). *NFKBIA* (deleted) and *KCNMA1* (amplified) were selected because they are novel and potentially interesting aberrations in CRC (Sections 5.2.3.5.2 and 5.2.3.5.3.5).

5.2.3.6.1 qPCR design

Pre-designed TaqMan copy number assays were chosen from the ABI website (www.appliedbiosystems.com, accessed April, 2011). A specific assay was chosen for each of the 3 genes. Figure 5.8 - Figure 5.10 represent the locations of the *PARK2*, *KCNMA1* and *NFKBIA* probes that are available on the Agilent 180K array platform. Additionally, the locations of the ABI TaqMan probes that were selected for the confirmatory qPCR experiments are shown. As described in Section 2.2.9.2, TaqMan copy number assays require a reference assay for relative copy number quantification. Two pre-designed and optimised copy number reference assays were available from ABI (*TERT* and *RNAP*). In order for the TaqMan copy number assays to be accurate, the reference probes should align to a diploid area of the genome. The genomic ploidy status of *PARK2*, *KCNMA1*, *NFKBIA* and the 2 reference assays according to aCGH results are shown in Table 5.10.

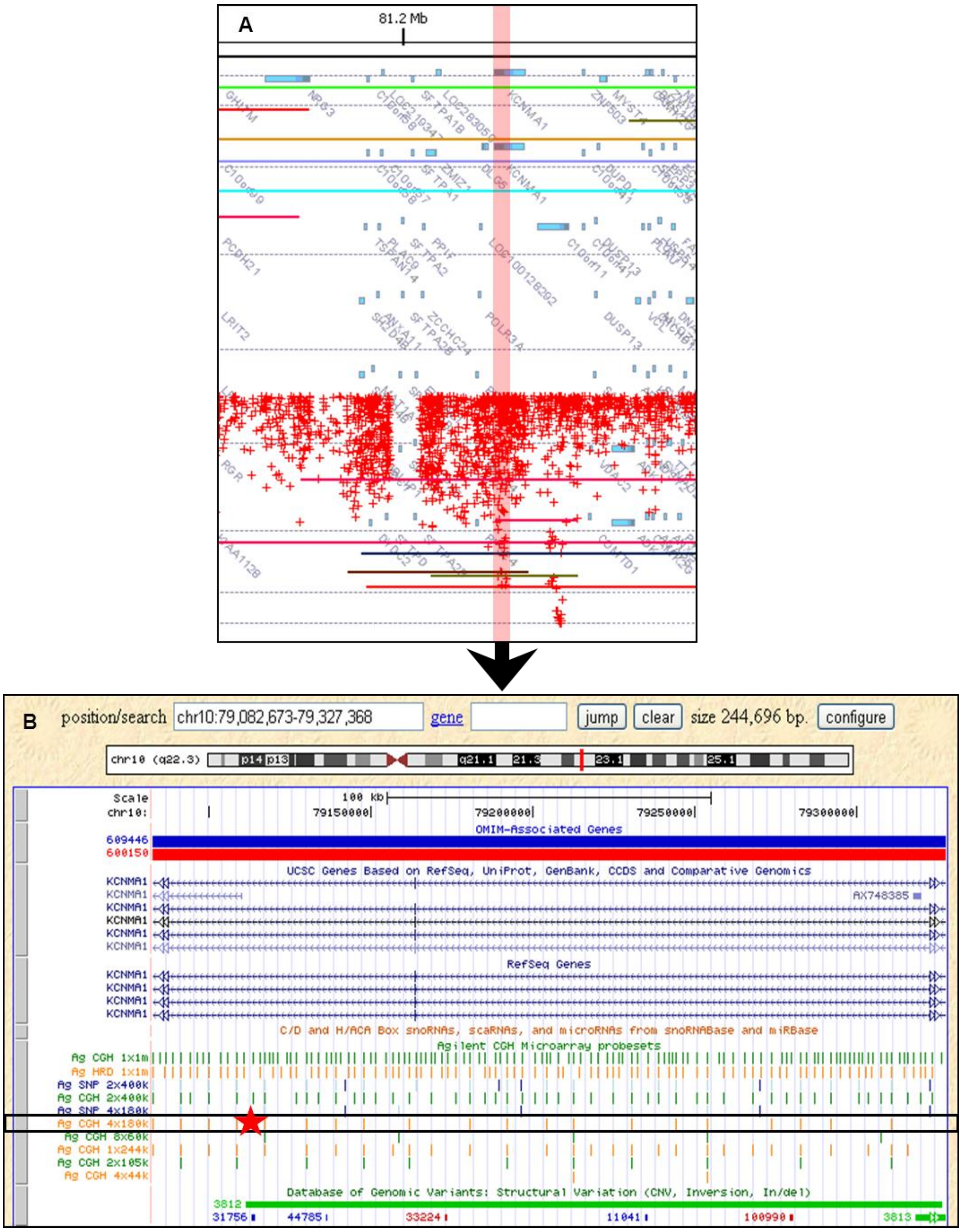


position/search chr6:162,607,448-162,652,882

size 45,435 bp.

Figure 5.8 ***PARK2*** **FMCR area and TaqMan copy number assay**

The FMCR area of *PARK2* on the genomic workbench software (Hg18). The aberrations appear as coloured lines, the green dots present deleted probes and the transparent green box represents the area shown in panel B. B) *PARK2* from the UCSC genome browser (Hg19) showing the Agilent probe locations from different platforms. The 180K array probes are marked by the box. The red star shows the approximate location of the TaqMan copy number probe. The CNVs located in the area are also shown.



B position/search chr10:79,082,673-79,327,368 gene jump clear size 244,696 bp. configure

Figure 5.9 ***KCNMA1*** **FMCR area and TaqMan copy number assay**

A) The FMCR area of *KCNMA1* on the genomic workbench software (Hg18). The aberrations appear as coloured lines, the red dots present amplified probes and the transparent red box represents the area shown in panel B. B) *KCNMA1* from the UCSC genome browser (Hg19) showing the Agilent probes locations from different platforms. The 180K array probes are marked by the box. The red star shows the approximate location of the TaqMan copy number probe. The CNVs located in the area are also shown.

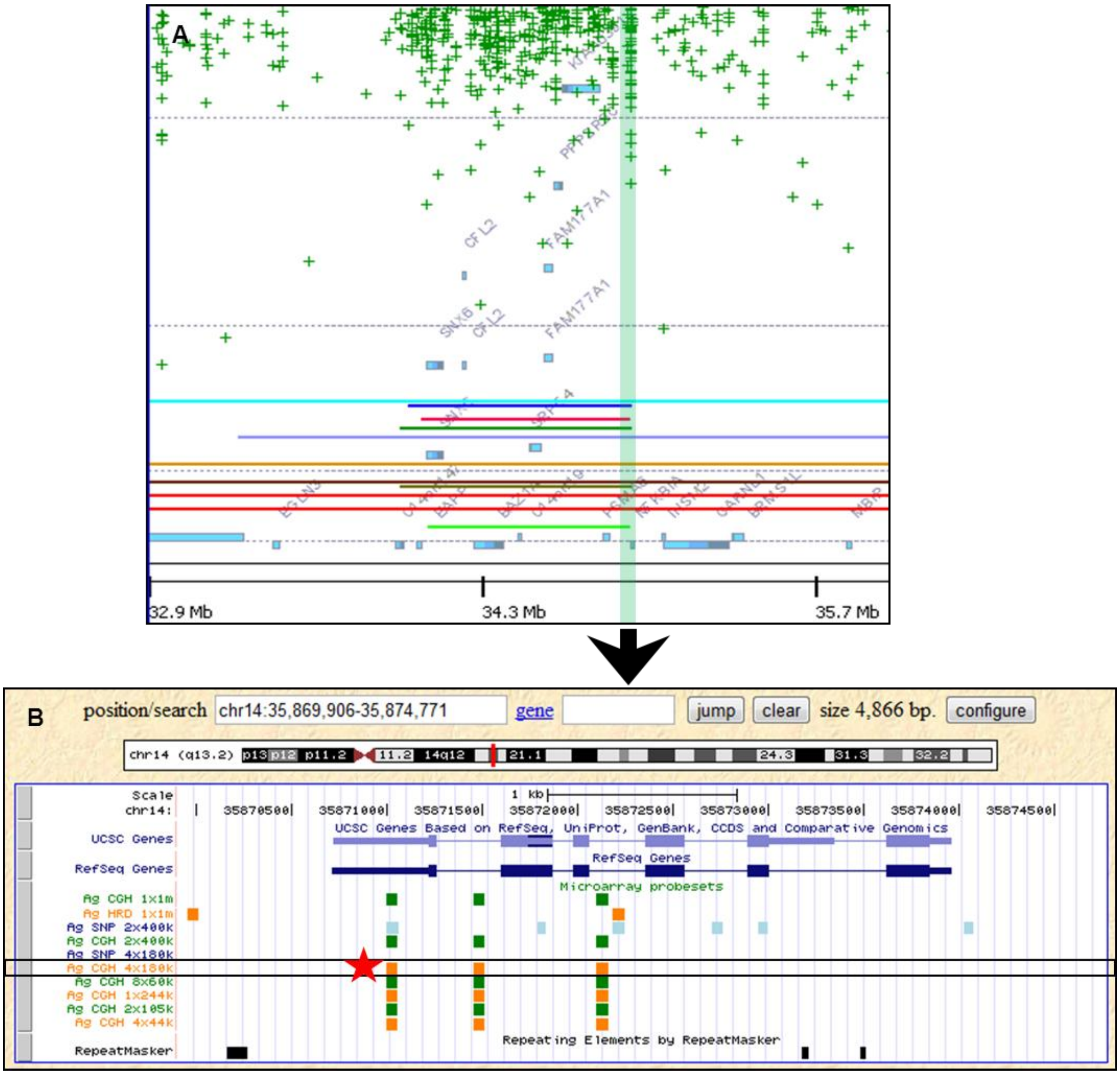


Figure 5.10 ***NFKBIA* FMCR area and TaqMan copy number assay**

A) The FMCR area of *NFKBIA* on the genomic workbench software (Hg18). The aberrations appear as coloured lines, the green dots present deleted probes and the transparent green box represents the area shown in panel B. B) *NFKBIA* from the UCSC genome browser (Hg19) showing the Agilent probes locations from different platforms. The 180K array probes are marked by the box. The red star shows the approximate location of the TaqMan copy number probe.

Table 5.10 Genomic status of *PARK2*, *NFKBIA*, *KCNMA1*, *TERT* & *RPPH1*

	Target Genes			Reference genes	
Sample	***PARK2***	***NFKBIA***	***KCNMA1***	***TERT***	***RPPH1***
T045		🟩	🟥	🟥	🟩
T053				🟥	
T079		🟩	🟥		🟥
T080		🟩		🟥	🟥
T088		🟩			
T090	🟩				🟩
T097	🟩				
T098		🟩			🟥
T107		🟩	🟥	🟥	
T114	🟩	🟩			
T142	🟩		🟥	🟥	
T158		🟩		🟥	🟩
T203		🟩		🟥	
T206	🟩		🟥		
T208			🟥		
T213			🟥		
T214		🟩			🟥
T221	🟩				
T223	🟩				🟥
T244		🟩			
T249	🟩	🟩		🟥	🟥
T828	🟩	🟩		🟥	🟩
T863			🟥		
T1350	🟩		🟥	🟥	

Table 5.10 depicts the samples with *PARK2* and *NFKBIA* deletions or *KCNMA1* amplifications. The table also summarises the genomic ploidy status of the reference genes *TERT* and *RPPH1* according to the aCGH results in the same CRC cases. Green, red and white boxes represent deletion, amplification and normal status respectively.

5.2.3.6.2 Assessment of *PARK2* deletion by qPCR

In total, 10 tumour DNA samples (22.2%) had deletions across the *PARK2* gene area (Table 5.10). T249 and T828 could not be confirmed because the genomic status of both *TERT* and *RPPH1* (the reference assays) was not diploid (Table 5.10). Also, T90 could not be confirmed, because its focal deletion was outside the area covered by the TaqMan probe. For the rest of the deleted samples, the genomic status of either the reference assay *TERT* or *RPPH1* was diploid. Copy number TaqMan assays steps were carried out according to the manufacturer's recommendations as described in Section 2.2.9.2.

Tumour DNA samples that required confirmation were T097, T114, T206, T142, T221, T223 and T1350. qPCR was performed on each tumour DNA sample and its normal control DNA pair in 4 replicates, and the findings are reported in Figure 5.11. *PARK2* deletions were confirmed in all of the tumour DNA samples except for T97 and T1350 (Figure 5.11). Table 5.11 summarises the calculated copy number ranges and their equivalent estimated copy number according to the Copy Caller software v1.0. Both T097 and T1350 calculated copy number was 1.67 and N097 and N1350 calculated copy numbers were 2.00, thus the results indicate a clear reduction in the copy number. However, as their calculated copy number was >1.5, the estimated copy number was 2. A closer look at the aCGH results of *PARK2* area in T097 and T1350 could explain the qPCR results (Figure 5.12). As shown in the Figure, T097 and T1350 clear focal deletions targeting *PARK2*. However, the deletions seem to be of low amplitude (indicated by low $\log_2$ratio) in comparison, for example, to T142.

Finally, all normal samples had a calculated copy number value very close to 2.0, except N142 which had a calculated copy number of 2.41. This might be caused by the CNVs in the area as shown above in Figure 5.8.

Table 5.11 Calculated and estimated copy numbers

Copy number	
Calculated	**Estimated**
<0.5	0
≥0.5-<1.5	1
≥1.5-<2.5	2
≥2.5-<3.5	3

Table 5.11 summarises the calculated copy number ranges and their equivalent estimated copy number according to the copy caller software.

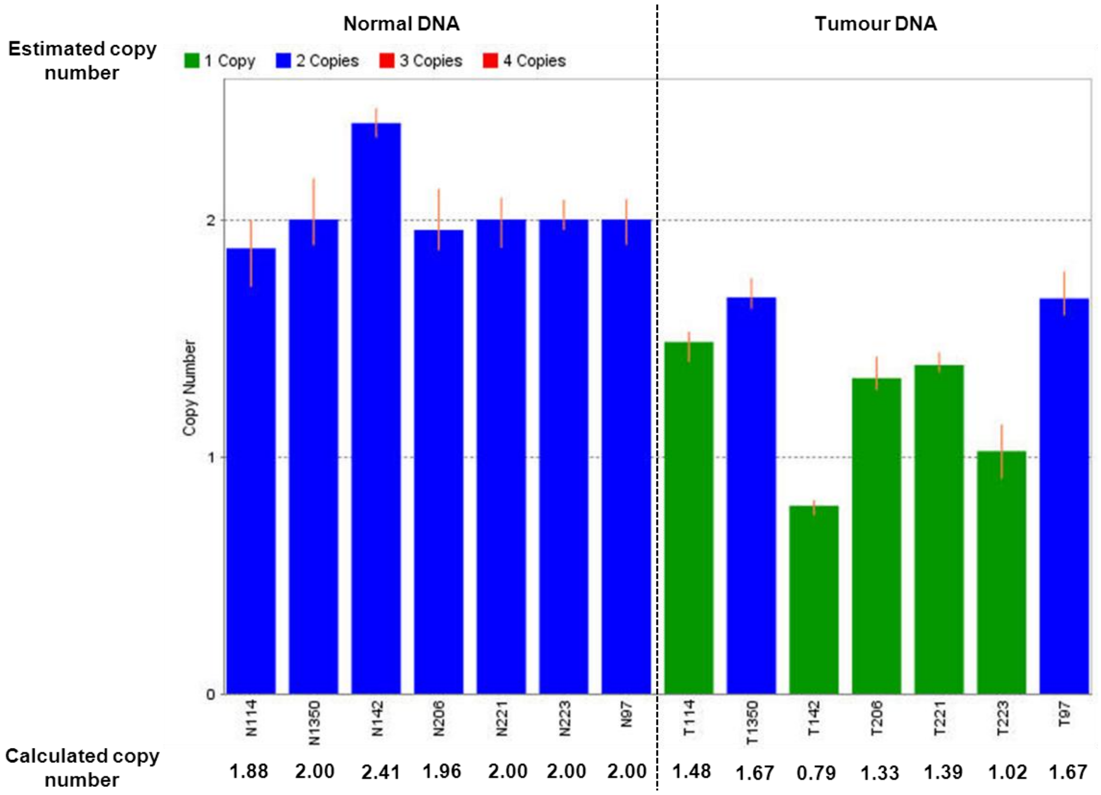


Figure 5.11 *PARK2* **QPCR results**

The Bars reflect the calculated copy number of the samples. The error bars represent the standard deviation based of 4 replicates within the same experiment. The colours of the bars reflect the estimated copy number.

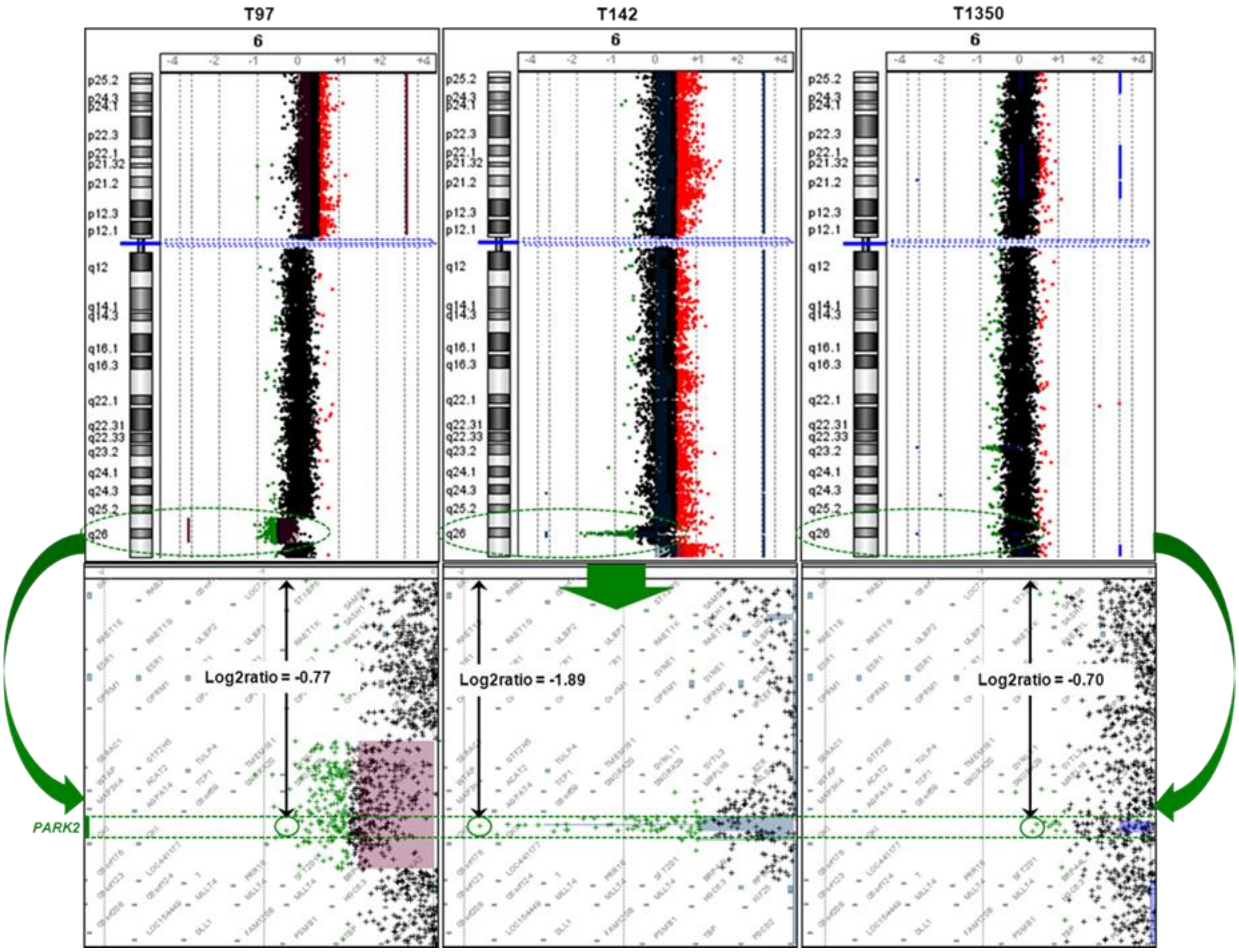


Figure 5.12 T097, T142 and T1350 *PARK2* aCGH results

Chromosome 6 aCGH data from T097, T142 and T1350, probes coloured in black represent genomic regions with a $\log_2$ratio around zero, probes coloured in red and green represent amplifications ($\log_2$ratio ≥ 0.5) and deletions ($\log_2$ratio ≤ -0.5) respectively. Closer look at *PARK2* show focal deletions with variable amplitudes (reflected by the $\log_2$ratio values).

5.2.3.6.3 Assessment of *KCNMA1* amplification by qPCR

In total, 9 tumour DNA samples (20.0%) had amplifications in *KCNMA1* (Table 5.10). T45 could not be confirmed because the genomic status of both *TERT* and *RPPH1* (the reference assays) was not diploid. Unfortunately, T107 and T863 could not be confirmed due to insufficient amounts of remaining DNA. For the rest of the samples, T079, T142, T206, T208, T213 and T1350, qPCR tests were performed as described. Paired normal DNA samples were also tested. The results are summarised in Figure 5.13. As shown in the Figure, qPCR confirmed *KCNMA1* amplifications in all of the tested tumour DNA samples. T208 calculated copy number (15.58) was very high, which is reflected in the aCGH results showing very high amplitude amplification (Figure 5.14).

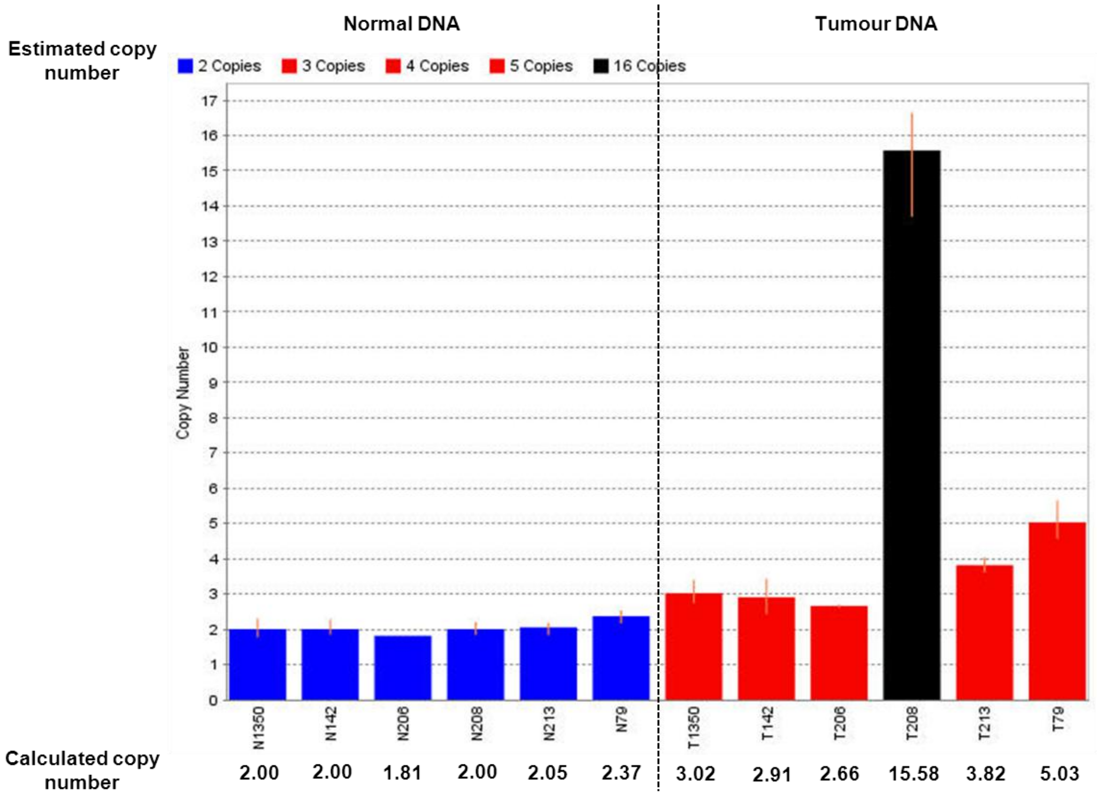


Figure 5.13 KCNMA1 qPCR results

The Bars reflect the calculated copy number of the samples. The error bars represent the standard deviation based of 4 replicates within the same experiment. The colours of the bars reflect the estimated copy number.

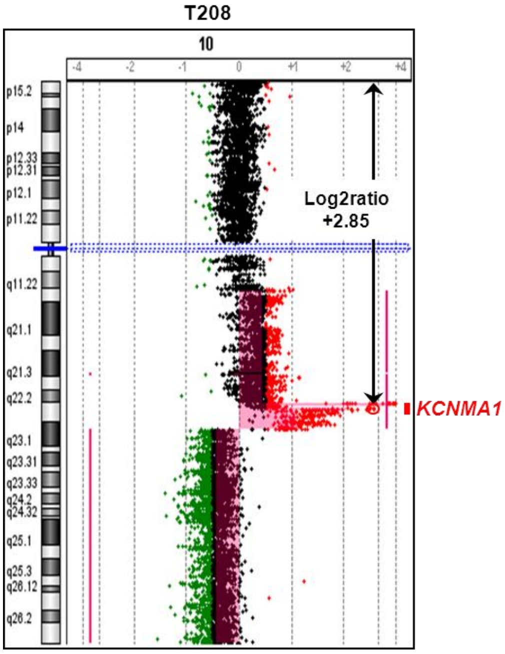


Figure 5.14 T208 *KCNMA1* aCGH results

Chromosome 10 aCGH data from T208, probes coloured in black represent genomic regions with a $\log_2$ratio around zero, probes coloured in red and green represent amplifications ($\log_2$ratio ≥ 0.5) and deletions ($\log_2$ratio ≤ -0.5) respectively. Very high $\log_2$ratio in *KCNMA1* area reflecting high amplitude amplification.

5.2.3.6.4 Assessment of *NFKBIA* deletion by qPCR

In total, 13 samples had deletions across the *NFKBIA* area. However, only 7 were available for confirmation. T045, T080, T158, T249 and T828 could not be confirmed because the genomic status of both *TERT* and *RPPH1* (the reference assays) was not diploid (Table 5.10). Moreover, T107 could not be confirmed due to insufficient amounts of remaining DNA. For the rest of the samples, T079, T088, T098, T114, T203, T214 and T244, qPCR testing was performed as described, earlier and the results of the tumour and normal DNA pairs are summarised in Figure 5.15. As shown in the figure, *NFKBIA* deletions were confirmed in all of the samples, with the exception of T244 (Figure 5.15). Close inspection of the aCGH data representing the *NFKBIA* area in T244 could potentially explain the results (Figure 5.16). As shown in the Figure, T244 had a focal deletion targeting *NFKBIA*. However, the deletion was of very low amplitude (indicated by a low $\log_2$ratio) in comparison to, for example T203.

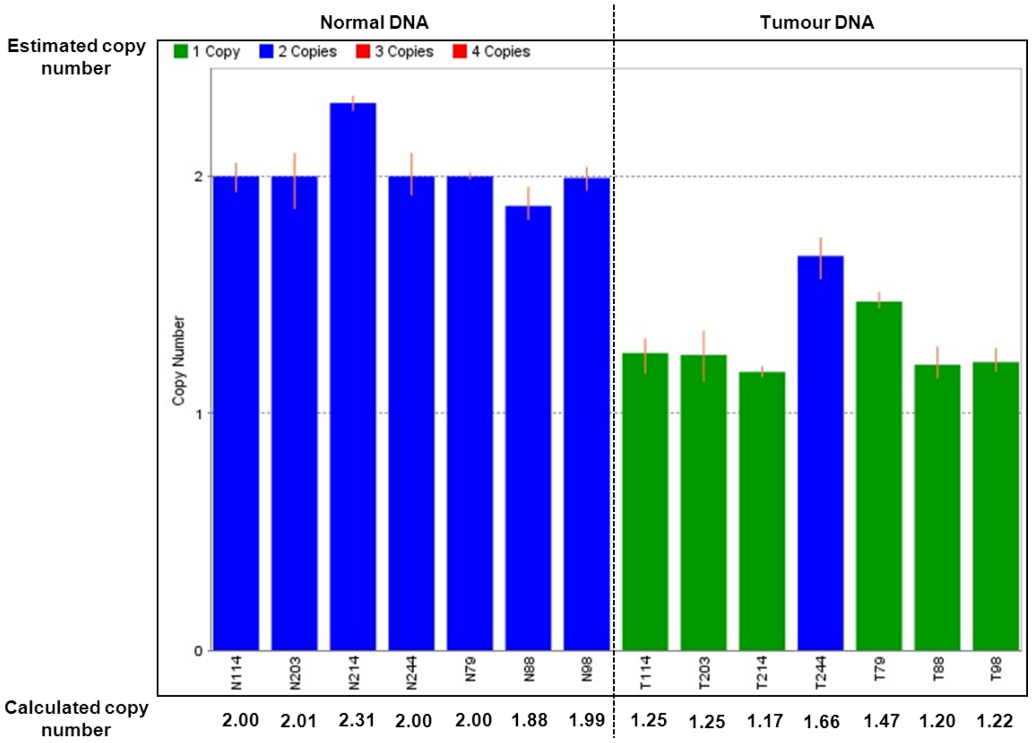


Figure 5.15 ***NFKBIA* QPCR results**

The Bars reflect the calculated copy number of the samples. The error bars represent the standard deviation based of 4 replicates within the same experiment. The colours of the bars reflect the estimated copy number.

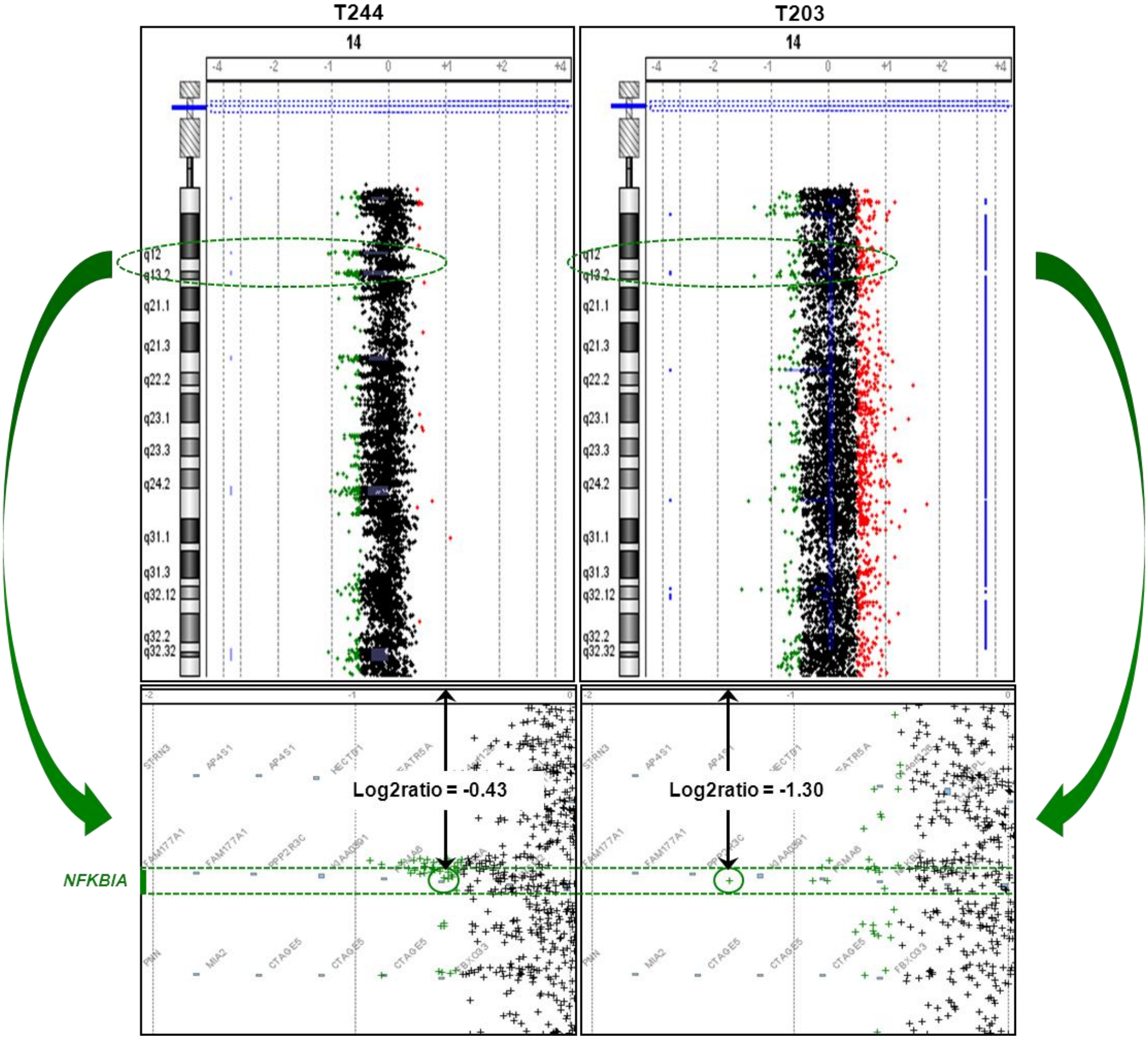


Figure 5.16 T244 and T203 *NFKBIA* aCGH results

Chromosome 14 aCGH data from T244 and T203, probes coloured in black represent genomic regions with a $\log_2$ratio around zero, probes coloured in red and green represent amplifications ($\log_2$ratio ≥ 0.5) and deletions ($\log_2$ratio ≤ -0.5) respectively. Closer look at *NFKBIA* show focal deletions with variable amplitudes (reflected by the $\log_2$ratio values).

5.2.3.7 Candidate FMCR and survival

In order to begin to assess the clinical significance of the above candidate deleted and amplified FMCR, a survival analysis was performed by Dr Angela Cox. This preliminary analysis was only performed for chromosomally unstable cases from the Sheffield population sample set (n=33) as the survival follow up data was not available for the cases from the Sheffield Royal Hallamshire Tissue bank sample set

(Section 2.1.3.4). The analysis suggested an association between *KCNMA1* amplification and poor prognosis (p=0.05) (Figure 5.17). There was no relationship between any of the other candidate FMCR and survival (p>0.05).

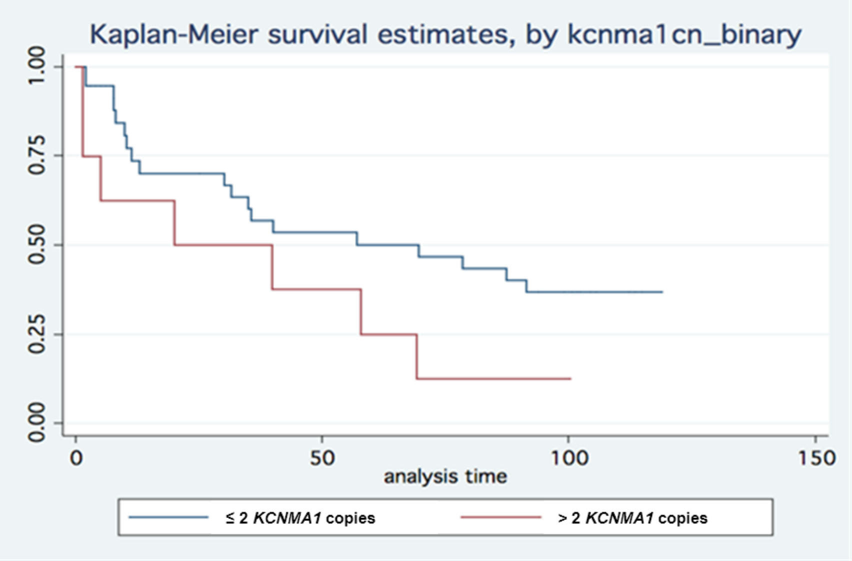


Figure 5.17 ***KCNMA1* amplification and survival**

Kaplan-Meier survival analysis estimated a worse survival of CRC cases with *KCNMA1* amplification (red line) in comparison to CRC cases without *KCNMA1* amplification (blue line).

It was noticed that the aCGH data showed that 2 of the 7 cases with *KCNMA1* amplification were focal and of high amplitude (T079 and T208) (Figure 5.18). The estimated copy number for these samples based on the aCGH results was >5 and this was quantitatively confirmed by qPCR (Figure 5.13). These 2 patients were diagnosed with CRC before 55 years of age and both died within 5 months of diagnosis due to CRC progression (Table 5.12). None of the remaining cases with survival information had a similar age of onset (<55 years) combined with a survival of less than 5 months.

Table 5.12 **CA079 and CA208 *KCNMA1* amplification and survival**

	***KCNMA1* amplification**			
	aCGH log2ratio*	**QPCR ****	**Age at diagnosis**	**Survival (months)**
CA79	1.67	5	54	5
CA208	2.85	16	51	1

* Maximum log2ratio value in the *KCNMA1* area.

** Estimated copy number.

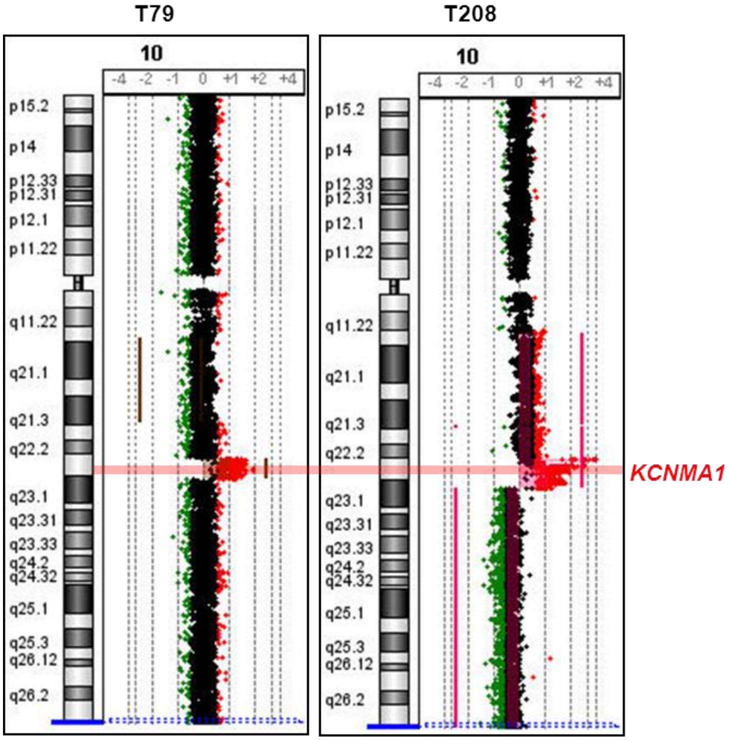


Figure 5.18 T079 and T208 focal *KCNMA1* amplification

Chromosome 10 aCGH data from T079 and T208, probes coloured in black represent genomic regions with a $\log_2$ratio around zero, probes coloured in red and green represent amplifications ($\log_2$ratio ≥ 0.5) and deletions ($\log_2$ratio ≤ -0.5) respectively. Very high $\log_2$ratio in *KCNMA1* area reflecting high amplitude amplification.

5.2.3.8 MicroRNA aberrations and *KCNMA1* and *NFKBIA*

The term microRNA (miRNA) refers to a group of small noncoding RNAs that regulate gene expression (Cummins *et al.* 2006). miRNAs were recently shown to play an important role in cancer development by altering the expression of both oncogenes and TSG (Schee *et al.* 2010). Several microRNAs are implicated in the development and progression of CRC (Arndt *et al.* 2009, Balaguer *et al.* 2010, Liu *et al.* 2010, Liu and Chen 2010). Deregulation of miRNA expression in CRC can result from altered methylation or chromosomal aberrations affecting miRNA genes (Diosdado *et al.* 2009, Balaguer *et al.* 2010, Liu and Chen 2010).

5.2.3.8.1 *KCNMA1* and microRNA211

Recently, miR211 was shown to target and down-regulate *KCNMA1* expression (Mazar *et al.* 2010). Under-expression of miR211 was shown to play a role in melanoma development and was directly associated with overexpression of *KCNMA1* (Mazar *et al.* 2010). MiR211 gene (*MIR211*) is located on the long arm of chromosome 15 and it was found to be deleted in ~24.4% (n=11) of the investigated CRC cases. Table 5.13 summarises the CRC cases with *KCNMA1* amplifications and/or *MIR211* deletions. As shown in the table, *KCNMA1* amplifications and *MIR211* deletions show a statistically significant mutual exclusive relationship (Spearman correlation test, p-value = 7.3x10^{-5}, r$_s$ = -0.80).

Table 5.13 *KCNMA1* amplifications and *MIR211* deletions

Sample	*KCNMA1* Amp	*MIR211* Del
T045	🟥	
T053		🟩
T079	🟥	🟩
T088		🟩
T097		🟩
T098		🟩
T107	🟥	
T142	🟥	
T206	🟥	
T208	🟥	
T213	🟥	
T214		🟩
T244		🟩
T249		🟩
T158		🟩
T741		🟩
T863	🟥	🟩
T1350	🟥	

Table 5.13 compares the CRC samples with *KCNMA1* amplifications (red) and/or *MIR211* deletions (green). The defects were shown to have a statistically significant inverse correlation (Spearman correlation test, p-value = 7.3x10^{-5}, r$_s$ = -0.80).

5.2.3.8.2 *NFKBIA* and miR301a

Recently, miR301a was reported to down-regulate NF-κB repressing factor (*NKRF*) resulting in NF-κB activation (Lu *et al.* 2011). The miR301a gene (*MIR301A*) is located on the long arm of chromosome 17 and it was found to be amplified in 22.2% (n=10) of the investigated CRC cases. Table 5.14 summarises the CRC cases with *NFKBIA* deletions and/or *MIR301A* amplifications. As shown in the table, *NFKBIA* deletions and *MIR301A* amplifications show a statistically significant mutual exclusive relationship (Spearman correlation test, p-value = 2.3×10^{-4}, r_s = -0.73)

Table 5.14 ***NFKBIA*** **deletions and** ***MIR301A*** **amplifications**

Sample	*NFKBIA* Del	*MIR301A* Amp
T045	🟩	
T079	🟩	
T080	🟩	
T088	🟩	
T098	🟩	
T107	🟩	🟥
T114	🟩	
T142		🟥
T153		🟥
T158	🟩	
T203	🟩	🟥
T206		🟥
T214	🟩	
T221		🟥
T244	🟩	
T249	🟩	
T741		🟥
T828	🟩	🟥
T863		🟥
T1350		🟥

Table 5.14 compares the CRC samples with *NFKBIA* deletions (green) and/or MIR301A amplifications (red). The defects were shown to have a statistically significant inverse correlation (Spearman correlation test, p-value = 2.3x10-4, rs = -0.73).

5.3 Discussion

As described previously, CIN is the most common form of genomic instability in CRC and is characterised by gains and losses of chromosomal segments or whole chromosomes (Sheffer *et al.* 2009, Migliore *et al.* 2011). CIN can drive cancer development by targeting TSG and oncogenes, many of which have not yet been identified (Brosens *et al.* 2010). Recently, the use of high resolution techniques such as SNP arrays and array CGH has prompted the identification of novel cancer driver genes in CRC and other cancer types (Bredel *et al.* 2010, Veeriah *et al.* 2010). As detailed, CIN was the main form of genomic instability found in our cohort (90.1%, n=48), and the aim of the work described in this chapter was to identify novel CRC driver genes affected by recurrent focal chromosomal aberrations.

5.3.1 Common aberration analysis

Statistically significant recurrent CNA were detected using the COCA algorithm. An in-house algorithm was then used to define FMCR based on the COCA output. FMCR were stringently defined as aberrant regions smaller than 3Mb in size, occurring in more than 10% of the cases with at least 2 SDE and a COCA score >2 (p-value = 0.01). Overall, 64 deleted and 32 amplified FMCR were identified according to these criteria. These focal aberrations were validated by comparison to recently published data in CRC from 4 different studies (Martin *et al.* 2007, Leary *et al.* 2008, Andersen *et al.* 2010, Brosens *et al.* 2010). Our FMCR have shown more overlap with the published focal aberrations when compared to the other 4 studies (Section 5.2.3.2). The overlapped regions helped in further delimiting the identified FMCR affecting well known cancer genes.

To further validate our FMCR, they were compared with the lists of the 20 most significant and recurrent focal deletions and amplifications identified across 26 different types of cancer including CRC (Section 5.2.3.3) (Beroukhim *et al.* 2010). A 25% (n=5) and 15% (n=3) overlap was identified between our deleted and amplified FMCR respectively. In general, comparison with the literature demonstrated the ability of the analyses methods and algorithms we have used to identify some the most commonly reported focal aberrations in CRC and other tumour types.

Pathway analysis using the online database DAVID revealed 3 significantly enriched cancer related pathways amongst the genes affected by the FMCR. For deleted FMCR genes, the most significantly enriched cancer-related signalling pathways were apoptosis and P53. Ten apoptotic genes were commonly deleted in our samples. Seven of these genes (*CASP3*, *CASP8*, *CASP10*, *NFKBIA*, *CAPN1*, *BAD*, *TNF* and *TNFRSF1A*) are pro-apoptotic and 3 are anti-apoptotic (*PIK3R2*, *PIK3R5* and *CFLAR*). It is noteworthy that the deleted anti-apoptotic gene, *CFLAR* occurs within the same FMCR as the pro-apoptotic genes *CASP8* and *CASP10*. Moreover, *PIK3R5* is located 1.2Mb downstream of *TP53* and 81% (n=17) of the deletions are common between the 2 genes. For the P53 signalling pathway, all the deleted genes (n=5) are known to have anti-tumourigenetic activities. It is notable that *TP53* was also deleted in 37.8% (n=17) of the cases, but it was not identified within an FMCR. These results indicate that the deleted FMCR might play an important role in tumour survival through disabling apoptosis and/or the P53 signalling pathway.

The most significantly enriched pathway for the amplified FMCR genes was the oncogenic MAPK pathway, with 9 genes being commonly affected. Seven of these genes (*CACNA1H*, *FGF23*, *FGF6*, *FGFR1*, *MAPKAPK2*, *ELK4* and *MYC*), are known or predicted to have oncogenic activities by promoting tumour growth and survival. It is noteworthy that one of the amplified anti-survival genes (*DUSP8*) occurred within the same FMCR with the oncogenic growth promoting gene *IGF2*. These results confirm that focal deletions and amplifications specifically target tumour suppressor and oncogenic pathways respectively.

5.3.2 Candidate CRC driver genes

In total, the identified FMCR contained 1002 genes. In order to identify a set of candidate genes, the FMCR were further shortlisted by increased stringency of their definition. Candidate FMCR were selected to have a maximum of 12 genes and they should be defined by at least 4 SDE. Based on the added criteria, shortlists of 17 deletions and 9 amplifications were identified.

The 17 candidate deleted FMCR harboured 71 genes, 16 (24.6%) of which are thought to be cancer-related. Consistently, 6 of these genes (*FHIT*, *TMSL3*, *PARK2*, *A2BP1*, *MACROD2* and *MAP2K4*) were recently reported to be deleted in CRC and other cancers (Leary *et al.* 2008, Andersen *et al.* 2010, Beroukhim *et al.* 2010, Veeriah *et al.* 2010). The remaining genes (*GMCL1*, *MAD1*, *PCBP1*, *CASP8*, *CASP10*, *FOLH1*, *CDK2AP1*, *NFKBIA*, *IFNAR1* and *IFNAR2*) were not previously reported to be deleted in CRC and the literature suggests potential role as TSG. Nonetheless, *MAD1* and *CASP8* were previously reported to be under-expressed (*MAD1*) and inactivated by somatic mutations (*CASP8*) in CRC patients (Gunes *et al.* 2000, Kim *et al.* 2003). *PCBP1* was recently shown to be under-expressed in CRC cell lines (Wang *et al.* 2010). Our data support a tumour suppressor role of these genes in CRC and suggest deletions as a novel mechanism of functional inactivation. Our data also propose *GMCL1*, *FOLH1*, *CDK2AP1*, *NFKBIA*, *IFNAR1* and *IFNAR2* as novel candidate CRC TSG.

The 11 candidate amplified FMCR harboured 44 genes, 8 (18.2%) of which were cancer related with predicted oncogenic functions (*PRDM16*, *PAX7*, *TNS1*, *RPA3*, *FGFR1, KCNMA1, FGF23* and *FGF6*). *FGFR1*, *FGF23* and *FGF6* were previously reported to be amplified in CRC and other cancer types (Sheffer *et al.* 2009, Beroukhim *et al.* 2010). Based on the literature, the remaining genes are proposed as novel candidate CRC oncogenes.

Importantly, copy number TaqMan assays specific for the well-established deleted *PARK2* gene and the novel *NFKBIA* deletion and *KCNMA1* amplification confirmed the accuracy of the aCGH data.

5.3.3 Candidate FMCR and survival

Survival analysis suggested an association between *KCNMA1* high amplitude amplifications and poor prognosis (Table 5.12). In addition to confirming its oncogenic potential, these results also suggest that *KCNMA1* amplification could be used as a potential prognostic marker. However, the results are based on a small sample size and should be confirmed in a larger sample before firm conclusions can be drawn.

5.3.4 MicroRNA aberrations and *KCNMA1* and *NFKBIA*

Selective pressure in tumour cells usually requires one mechanism to achieve a certain tumorigenic advantage (Issa 2008, Rinkenbaugh and Baldwin 2011). This can be seen in known mutually exclusive relationships found in CRC, including *BRAF* V600E and *KRAS* mutations (Davies *et al.* 2002). *KRAS* and *BRAF* both regulate the MAPK signalling pathway (Nagasaka *et al.* 2004). Activating mutations in *KRAS* and the *BRAF* V600E mutation are known to constitutively activate the MAPK pathway, thus, their co-existence in the same tumour is extremely rare (Yuen *et al.* 2002, Kumar *et al.* 2009). Statistically significant inverse relationships were observed between *KCNMA1* amplifications and *MIR211* deletions and *NFKBIA* deletions and *MIR301A* amplifications. These inverse correlations support the role of miR211 in activating *KCNMA1* and the role of both NFKBIA and miR301A in regulating NF-κB. Moreover, these observations support the oncogenic role of *KCNMA1* amplifications and *NFKBIA* deletions.

In summary, our results supported the hypothesis that recurrent focal aberrations selectively target cancer-related genes. Moreover, focal deletions and amplifications were clearly established to affect known TSG and oncogenes respectively. Several novel candidate CRC driver genes were also proposed with potential prognostic and therapeutic value in the future.

6. *NFKBIA*, a Tumour Suppressor in CRC

6.1 Introduction

Nuclear factor of κ-light polypeptide chain (NF-κB) is a transcription factor that plays an important role in promoting cell proliferation and survival (Sethi *et al.* 2007, Bredel *et al.* 2010). NF-κB is comprised of dimers of 5 subunits: RelA (p65), RelB, c-Rel, p50 and p52 (Lu *et al.* 2011, Rinkenbaugh and Baldwin 2011). The most abundant form of NF-κB heterodimers is RelA-p50 (Kojima *et al.* 2004, Ma *et al.* 2011). Under basal conditions, NF-κB dimers remain inactive in the cytoplasm by the binding of inhibitors of κB (IκB) proteins (Karin *et al.* 2002).

NF-κB activation occurs through 2 main pathways; the canonical classical pathway and the alternative pathway (Inoue *et al.* 2007). Both pathways result in the translocation of NF-κB dimers into the nucleus and the subsequent transactivation of their target genes (Bonizzi and Karin 2004, Inoue *et al.* 2007). NF-κB targets include immunoregulatory, inflammatory, anti-apoptotic and cell proliferative genes (Wang *et al.* 1996, Karin *et al.* 2002, Pikarsky *et al.* 2004, Jeyasuria *et al.* 2011). Proinflammatory cytokines, such as the tumour necrosis factor (TNF) family are important activators of the NF-κB pathway (Wang *et al.* 1996, Karin *et al.* 2002). Aberrant activation of NF-κB is known to play an oncogenic role in the development of several cancer types including, pancreatic and prostate cancer, Hodgkin lymphoma, GBM and CRC (Krappmann *et al.* 1999, Lind *et al.* 2001, Suh *et al.* 2002, Kojima *et al.* 2004, Inoue *et al.* 2007, Meylan *et al.* 2009, Basseres *et al.* 2010, Bredel *et al.* 2010, Lu *et al.* 2011). Aberrant NF-κB activation is relevant for the prognosis and survival of cancer patients, in addition to treatment efficiency (Camp *et al.* 2004, Izzo *et al.* 2006, Scartozzi *et al.* 2007). Subsequently, the NF-κB pathway and its regulatory proteins are now considered as potential targets for novel cancer therapeutic strategies (Inoue *et al.* 2007, Rinkenbaugh and Baldwin 2011).

Somatic aberrations affecting NF-κB regulatory genes and microRNAs are responsible for the inappropriate activation of the pathway in several cancer types (Cabannes *et al.* 1999, Krappmann *et al.* 1999, Nishikori 2005, Lu *et al.* 2011). Inactivating mutations in *NFKBIA*, encoding NF-κB inhibitor alpha, were previously shown to play a role in Hodgkin lymphoma (Cabannes *et al.* 1999). Recently, heterozygous *NFKBIA* deletions were reported in ~25% of GBM cases (Bredel *et al.* 2010). Moreover, siRNA knockdown of *NFKBIA* was shown to result in increased levels of the activated phosphorylated RelA in lung cancer cell lines (Bivona *et al.* 2011).

NFKBIA was predicted to be the driver gene in one of the candidate deleted FMCR identified in this project (Section 5.2.3.5.2). The aCGH data and the $\log_2$ratio values for this deletion FMCR were of relatively low amplitude, suggesting a heterozygous deletion, which agrees with the findings reported in GBM (Bredel *et al.* 2010). Moreover, the frequency of the deletions was 28.9%, also similar to that in GBM (Table 5.8). Therefore, as proof of concept experiments, preliminary functional studies investigating the role of *NFKBIA* deletions in CRC development were performed. The two major aims of these experiments described in this chapter were as follows. Firstly, to examine NFKBIA protein expression levels in a panel of CRC cell lines, in particular to observe whether any CRC cell lines contained reduced levels of NFKBIA. Two GBM cell lines, U87 and SNB19, were included as controls, because NFKBIA is known to be underexpressed in GBM (Bredel *et al.* 2010). Finally, two non-tumour derived cell lines, MRC5-SV2 and HEK293, were also employed. Secondly, to investigate the effect of NFKBIA depletion on NF-κB activation and malignant behaviour in culture, by performing NFKBIA knockdown, using *NFKBIA*-siRNA, in CRC cell lines expressing high levels of NFKBIA. Subsequently, these cell lines were used to perform clonogenic assays.

6.2 Results

6.2.1 NFKBIA in CRC cell lines

Western blotting (Section 2.2.17) was performed on cellular extracts from 6 CRC cell lines, HCT116, CACO2, SW620, SW480, COLO205 and HT29, in addition to 2 GBM cell lines, U87 and SNB19, and 2 non-tumour derived cell lines, MRC5-SV2 and HEK293. These results are presented in Figure 6.1. Remarkably, NFKBIA protein was expressed at lower levels in 3 of the 6 CRC cell lines (CACO2, SW480 and COLO205). The other 3 cell lines (HCT116, SW620 and HT29) had similar or increased levels of NFKBIA compared to the non-tumour cell lines. SNB19 (GBM) had low levels of NFKBIA (similar to CACO2 and SW480) and U87 (GBM) had similar levels to HT29. These experiments were repeated 3 times using 2 independent cellular extracts.

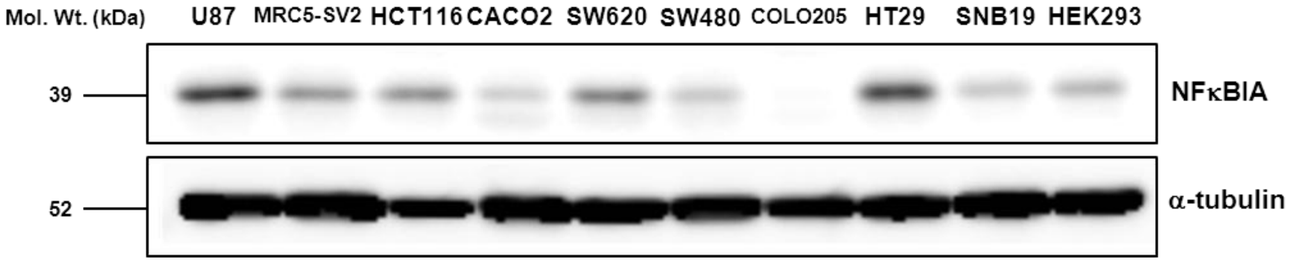


Figure 6.1 ***NFKBIA*** **protein expression**

Western blot showing NFKBIA protein levels in CRC, GBM and control cell lines. An alpha-tubulin antibody was used to confirm loading equal amounts of cellular extracts from each cell line.

6.2.2 NFKBIA and phosphorylated NF-κB

In order to examine the effect of *NFKBIA* deletions on NF-κB activation and tumour behaviour, *NFKBIA*-siRNAs were transfected into 3 CRC cell lines (HCT116, SW620 and HT29) that had relatively increased levels of NFKBIA. siRNA experiments were performed as described in Section 2.2.18. For each cell line, a negative control was performed by treating the cells with “negative” siRNA designed not to complement any known human mRNA sequence. Western blotting with NFKBIA and p-RELA antibodies was performed on cellular extracts taken 24 and 48 hr after transfection of CRC cell lines with *NFKBIA*-siRNA. Figure 6.2 summarises the Western blotting findings. As indicated, a marked reduction in NFKBIA expression was confirmed in HCT116 and SW620 cells. However, the effect appeared of short duration, because the NFKBIA levels were lower 24 hr after transfection when compared to 48 hr. Additionally, p-RELA levels were very low, but detectable, in all of the 3 cell lines regardless of siRNA treatment and NFKBIA expression. The Western blot was performed twice on 2 separate siRNA transfections.

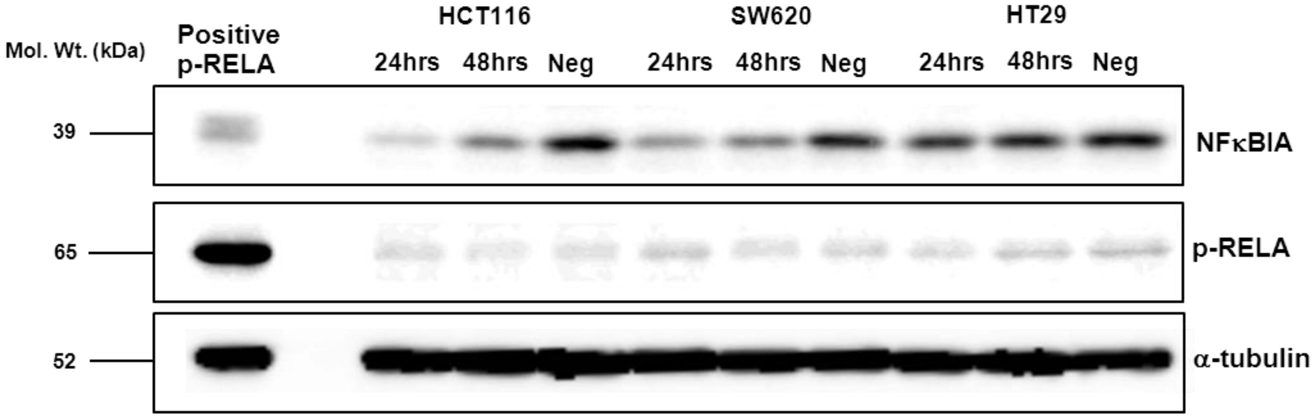


Figure 6.2 NFKBIA siRNA experiment

HCT116, SW620 and HT29 cells were transfected with *NFKBIA*-siRNA and “negative” siRNA (Neg). Protein extraction was performed 24hrs and 48hrs after *NFKBIA*- siRNA transfection. For the negative controls, protein extraction was performed 48hrs post transfection with “negative” siRNA. Western blotting was then performed using NFKBIA and p-RELA antibodies. p-RELA positive control extract (HeLa cells treated with TNF-α) was also included.

6.2.3 Effects of NFKBIA knockdown in cell culture

The effects of NFKBIA protein depletion were evaluated by observing tumour cell proliferation and colony formation in HCT116 and SW620 cells transfected with *NFKBIA*-siRNA and “negative” siRNA.

HCT116 and SW620 cell proliferation was measured via the MTT assay (Section 2.2.19). The MTT test was performed on 3 consecutive days following treatment of both cell lines with *NFKBIA*-siRNA and the “negative” control siRNA. The results are summarised in Figure 6.3.

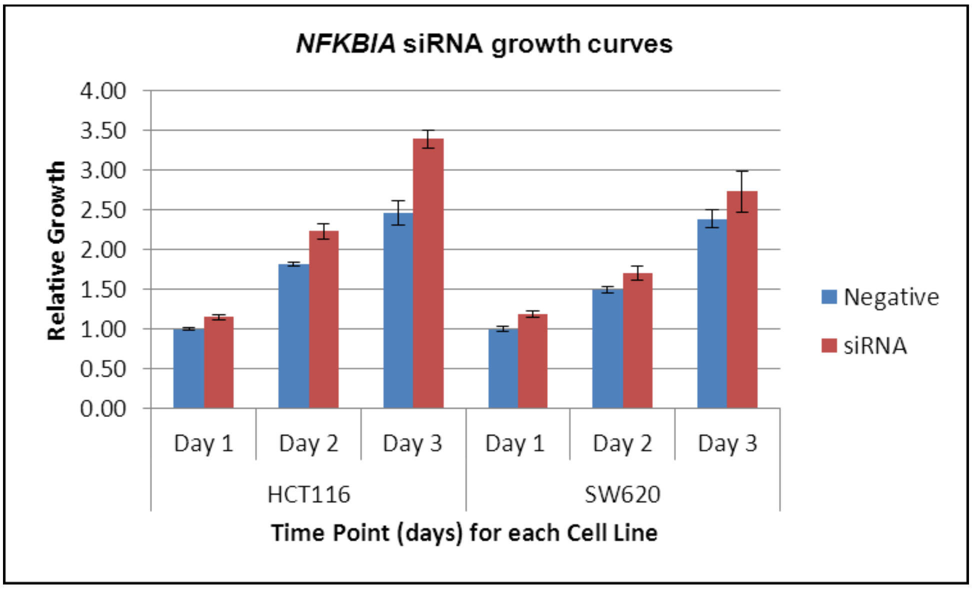


Figure 6.3 MTT assay results

Cell proliferation was inferred spectrophotometrically via the MTT assay. The growth levels were calculated relative to day 1 negative control for each cell line. The error bars represent the standard deviation of 4 replicates within the same experiment. The above results are based on a single experiment.

Colony forming ability of tumour cells in culture is known to reflect their malignancy (Bredel *et al.* 2010). The colony formation test was performed as described in Section 2.2.20. HCT116 and SW620 colonies were counted 9 and 13 days post siRNA transfection respectively. As for the MTT assay (above), “negative” siRNA controls were also included. Five replicates were performed for each cell line and treatment. Figure 6.4 shows the colony formation results for 2 replicates of HCT116 cells treated with *NFKBIA*- siRNA and “negative” siRNA. Figure 6.5 summarises the colony counts from the 5 replicates for both cell lines and treatments.

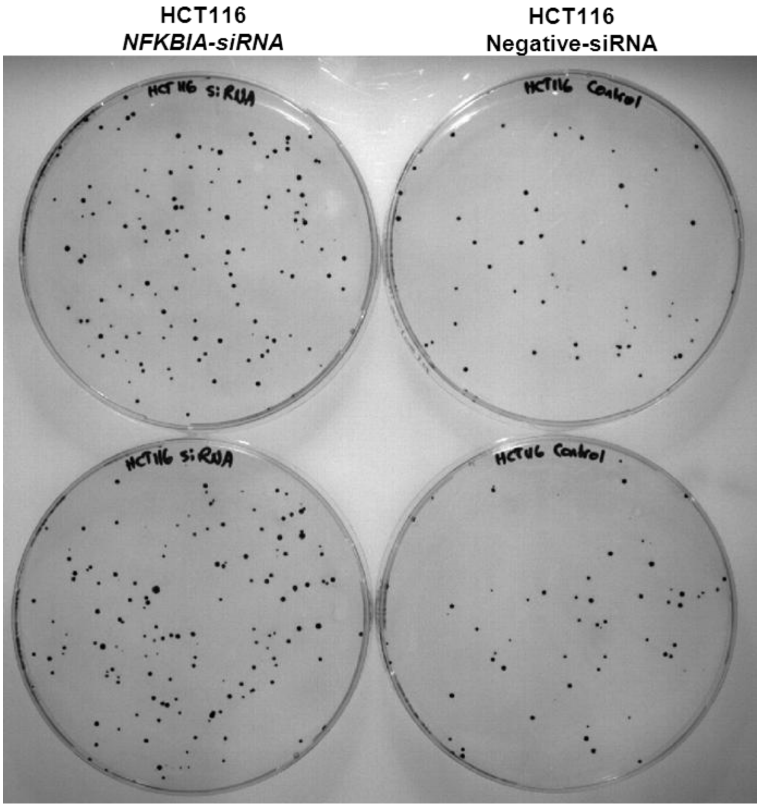


Figure 6.4 Colony formation test in HCT116 (2 replicates)

The 4 petri dishes represent 2 replicates for colony formation test for HCT116 cells treated with *NFKBIA*-specific siRNA and negative siRNA.

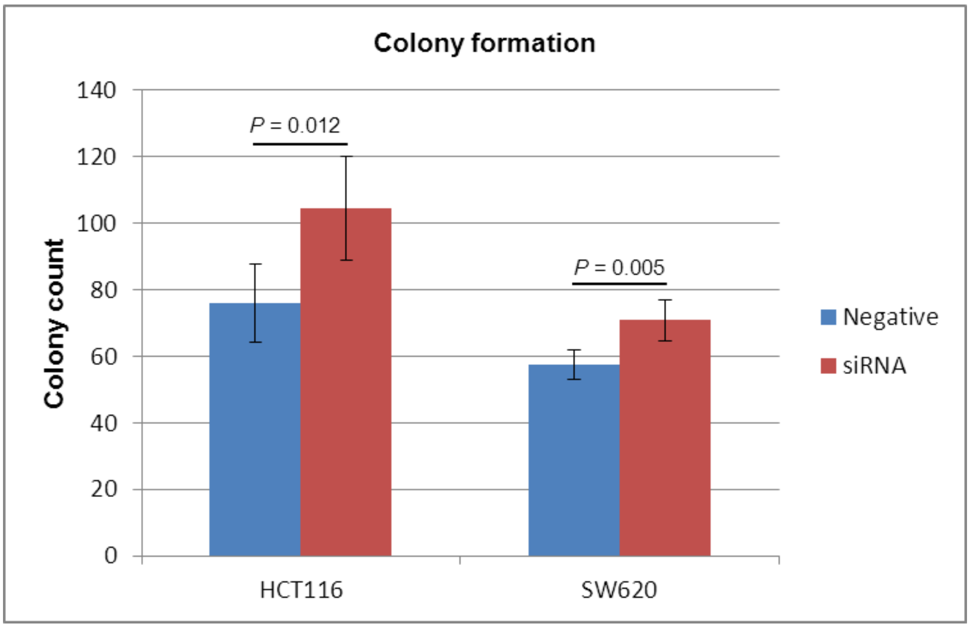


Figure 6.5 Colony formation test results

Colony counts for each cell line were based on 5 replicates. P-values were calculated using unpaired t-test (two tailed).

6.3 Discussion

The novel candidate CRC TSG *NFKBIA* was further investigated and its protein was found to be relatively under-expressed in 3 of the 6 CRC cell lines investigated. NFKBIA-siRNA transfection was used to deplete NFKBIA protein in the CRC cell lines with higher levels of NFKBIA (HCT116, SW620 and HT29). Lower levels of NFKBIA proteins were detected in HCT116 and SW620 cell lines after siRNA treatment (Figure 6.2). We were unable to determine whether the lower levels of NFKBIA protein affect the levels p-RELA, as previously suggested (Bivona *et al.* 2011). The levels of p-RELA were very low in all CRC cell lines tested regardless of NFKBIA levels. In contrast, p-RELA levels were higher in lung cancer cell lines, and an increase in level was observed with depletion of NFKBIA (Bivona *et al.* 2011).

Knockdown of NFKBIA was also investigated for any potential effect on CRC cell line malignant behaviour as measured by cell growth and colony formation. Following transfection with *NFKBIA*-siRNA and "negative" siRNA control, lower levels of NFKBIA increased cell proliferation and the ability to form colonies in both HCT116 and SW620. However, the effect seems to be more evident in HCT116, which could potentially be due to the lower levels of NFKBIA observed in HCT116 following siRNA transfection (Figure 6.2). These results agree with the effect of NFKBIA re-expression in GBM cell lines (Bredel *et al.* 2010). In their study, Bredel and colleagues achieved NFKBIA overexpression in 2 GBM cell lines which resulted in decrease in cell proliferation and the ability to form colonies in culture and increased sensitivity to the chemotherapeutic agent Temozolomide (Bredel *et al.* 2010). As a proof of concept, these experiments support the predicted tumour suppressor role of *NFKBIA* in CRC. Nevertheless, these represent preliminary results that need to be repeated and confirmed.

7. Discussion and Conclusions

7.1 *CASP8* inherited variants and CRC risk

Cancer in is considered a genetic disease, as both germline and somatic genetic factors play an essential part in its predisposition, initiation and development. CRC is considered the 2nd highest amongst common cancers, behind prostate, in terms of inherited susceptibility, with an estimated genetic contribution of ~35% (Lichtenstein *et al.* 2000). Highly and moderately penetrant genetic factors are rare in CRC and account for ~5-7% of the total CRC cases. The rest of the heritability is predicted to be caused by less penetrant but more common genetic factors (Broderick *et al.* 2007, Lascorz *et al.* 2010). Several low penetrant and common CRC risk variants were identified through genome wide and candidate genes association studies, however, so far they only account for ~6% of CRC heritability (Lascorz *et al.* 2010). Thus, most of the inherited susceptibility for CRC remains unknown.

Caspase 8 is an initiator caspase that plays an important role in activating apoptosis (Grenet *et al.* 1999). Loss of caspase 8 protein function through acquired mutations, deletions or silencing of *CASP8* gene has been shown to play a role in the development of several cancer types including CRC (Teitz *et al.* 2000, Kim *et al.* 2003, Pingoud-Meier *et al.* 2003b, Soung *et al.* 2005, Stupack *et al.* 2006). Moreover, *CASP8* inherited variants were shown to affect the risk of several cancer types including lung and breast cancer (Son *et al.* 2006, Cox *et al.* 2007). A somewhat controversial association between an inherited 6bp in/del variant in the *CASP8* promoter and CRC risk had been proposed at the beginning of this project (Sun *et al.* 2007, Haiman *et al.* 2008, Pittman *et al.* 2008). Moreover, preliminary data from our laboratory indicated a possible association between some *CASP8* SNPs and CRC risk (Curtin *et al.* manuscript in preparation). The aims of the work described in chapter 3 were to investigate the controversial role of the rs3834129 *CASP8* promoter variant and CRC risk, to identify novel coding *CASP8* variants that can affect CRC risk, and to develop an assay to investigate a possible role of *CASP8* CNV23598 in CRC and breast cancer predisposition.

7.1.1 Summary and conclusions

The *CASP8* promoter in/del rs3834129 was genotyped in 1193 CRC cases and 1388 controls. The cases and controls were from the UK (Sheffield, Leeds and Dundee) and USA (Utah). Association analysis was performed on the 4 sample sets separately and combined, and the results did not indicate any significant association between rs3834129 genotypes and CRC risk (p-value > 0.05). This conclusion was in agreement with the published data on the European and multi-ethnic American populations (Haiman *et al.* 2008, Pittman *et al.* 2008), however, it contradicted the original results showing association in the Chinese populations (Sun *et al.* 2007). The disparity between the original study in the Chinese population and our results (in addition to the results from the European and American populations) could be due to different effects of genetic and environmental modifier factors between the different populations, or due to the initial association being a false positive. Nevertheless, the rs3834129 data is being included in a more comprehensive analysis of SNPs in the *CASP8* gene region and their potential role in CRC risk (Curtin *et al.* manuscript in preparation). This study aims to fine map the region and examine any possible associations with CRC susceptibility in various subgroups of the cases. The results do suggest a possible association between *CASP8* SNPs and female early onset CRC (Curtin *et al.* manuscript in preparation).

The *CASP8* promoter region, exons, intron/exon boundaries and 3’ UTR were sequenced in 94 CRC cases from the Sheffield population sample set. The sequencing results revealed the presence of 6 (relatively rare and) novel variants with minimum allele frequencies of 0.005 to 0.01. To date, none of these variants have been previously reported. *In silico* analysis was performed and the results did not predict any functional impact of these variants on caspase 8. Nonetheless, the 115 bp deletion in *CASP8* exon 3 (c.1-8338Del115) and the variant in *CASP8* intron 3 (c.1-1982A>G) occurred within a region with a possible regulatory role assessed by the presence of DNase hypersensitivity clusters which could reflect the presence of transcriptional enhancers and promoters (Crawford *et al.* 2006). However, this region is expressed as part of the precursor of the non-predominantly expressed *CASP8* isoform-G which is not detected on protein level (Scaffidi *et al.* 1997). Therefore, none of the variants warranted further investigation at this stage. But, if the results of the fine mapping by Curtin and colleagues suggest that either of these two variants (c.1-8338Del115 and c.1-1982A>G) occur on CRC associated risk haplotypes, they may merit further investigation to examine possible effects on expression of *CASP8* isoform transcripts.

CASP8 CNV23598 was genotyped in 284 CRC cases and controls (Sheffield sample set) and 47 breast cancer cases and controls enriched for *CASP8* risk and protective haplotypes respectively. The initial genotyping results indicated differences in CNV23598 allele frequencies for CRC and breast cases and controls. However, the CRC control and case genotype frequencies were inconsistent with HWE, which might indicate genotyping errors. Several molecular techniques confirmed the genotyping results. However, they indicated the presence of multiple CNV23598 insertion copies in addition to differences from the published insertion sequence at both breakpoints. A few more CNVs with variable lengths were recently reported around the CNV23598 region which could explain some of the genotyping issues (http://projects.tcag.ca/variation/, accessed August, 2011). *CASP8* region is currently being fully investigated using a second generation sequencing platform in the 47 breast cancer cases and controls as part of a different project. The sequencing results might help in further elucidating the reasons behind the questionable results for CNV23598. In conclusion, our results do not provide evidence that *CASP8* inherited variants have any significant effect on CRC risk overall. Although more detailed investigation including subgroup analysis is underway. Finally, we have shown that CNV23598 has a very complex architecture that may only be resolved by advanced molecular techniques (such as 2nd generation sequencing).

7.2 Molecular classification of CRC

Sporadic CRC is highly heterogeneous from a molecular point was shown to develop through multiple pathways of genomic instabilities (MSI, CIMP and CIN), and somatic mutations in key driver genes (*APC*, *TP53*, *KRAS*, *BRAF* and *PIK3CA*) (Cheng *et al.* 2008, Derks *et al.* 2008). Molecular classification of sporadic CRC has potential applications in research and in the clinic. It can help in understanding the underlying mechanisms of CRC development, identify new molecular targets and pave the way for personalised cancer care (Choong and Tsafnat 2011, Pritchard and Grady 2011). One of the main drawbacks in the field is the lack of standardisation in investigating and defining the molecular characteristics, especially the global genomic instabilities CIMP and CIN (Cheng *et al.* 2008, Derks *et al.* 2008, Ogino and Goel 2008). The main aim of the work described in chapter 4 was to classify a cohort of matching CRC patients using a more standardised approach.

7.2.1 Summary and conclusions

A thorough molecular characterisation was performed on 53 CRC cases with matching tumour and normal DNA samples. MSI was defined with a panel of 5 mononucleotide microsatellite markers according to the revised NCI criteria for MSI testing (Bacher *et al.* 2004, Umar *et al.* 2004). CIMP was investigated using a semi-quantitative MS-MLPA technology based on a set of 7 CpG markers previously verified in 3 independent studies (Ogino *et al.* 2006a, Weisenberger *et al.* 2006, Ogino *et al.* 2007). CIN was investigated using a high resolution genome wide array CGH technology and defined as recently recommended by Cheng and colleagues

(Cheng *et al.* 2008). DNA Sequencing was used to identify the most common mutations in *APC*, *TP53*, *KRAS*, *BRAF* and *PIK3CA*. All the investigated samples were shown to be MSS. Three molecular subtypes were identified within the investigated cohort, chromosomally stable/CIMP-L subtype (~7.7%), chromosomally unstable/CIMP-L/N subtype (78.8%) and a novel chromosomally unstable/CIMP-H subtype (11.5%). The latter may be a novel subtype missed in previous studies due to technical limitations in defining CIN, or due to the association of CIMP-H with MSI-H which has an established inverse correlation with CIN (Goel *et al.* 2007, Cheng *et al.* 2008). The existence of CIN/CIMP-H/MSS as a distinct molecular subtype from CIN/CIMP-L/N/MSS is supported by the different patterns of broad common CNA and *BRAF*/*KRAS* mutations. Nevertheless, these results need to be confirmed independently in a larger cohort.

7.3 Focal chromosomal aberrations and CRC

Chromosomal instability is the most common form of genomic instability in CRC and it is known to drive CRC development by affecting TSG and oncogenes (Martin *et al.* 2007, Brosens *et al.* 2010). CIN is usually characterised by a large number of chromosomal aberrations that affect a wide range of genes (Rajagopalan *et al.* 2003). Most of the key driver genes affected by chromosomal abnormalities are yet to be identified (Brosens *et al.* 2010). The aim of the work described in chapter 5 was to investigate the genes affected by common focal chromosomal aberrations in order to identify novel candidate driver genes.

7.3.1 Summary and conclusions

Array CGH was successfully performed on 53 CRC cases. Common aberration analysis was performed on the 45 cases that had high values for QC metrics and showed chromosomal instability. Chromosomal aberrations were identified using strict algorithms and definitions. Frequencies of the most common broad CNA including amplifications within chromosomes 20q, 13, 8q, 7 and X and deletions within chromosomes 18, 8p and 17p strongly agree with published data (Hermsen *et al.* 2002, Lassmann *et al.* 2007, Poulogiannis *et al.* 2010a). Context corrected common aberration analysis combined with an in-house developed algorithm were used to identify FMCR. Sixty-four deleted and 32 amplified FMCR were identified. Deleted FMCR contained established TSG such as *SMAD4* and *CDKN2C*, and amplified FMCR contained established oncogenes such as *MYC* and *FGFR1*. Comparison with previously published data in CRC and other cancer types (Martin *et al.* 2007, Leary *et al.* 2008, Andersen *et al.* 2010, Beroukhim *et al.* 2010, Brosens *et al.* 2010) have shown enrichment within our FMCR of some of the most commonly reported focal aberrations which supported the efficiency of the used algorithms.

Pathway analysis using the online database DAVID revealed that genes within cancer related pathways were commonly affected by the identified FMCR. The anti-survival apoptosis and P53 pathways were affected by deleted FMCR and the oncogenic MAPK pathway was affected by amplified FMCR. The enrichment of apoptotic genes within the deleted FMCR supports the correlation between genomic instability and defective apoptosis (Zhivotovsky and Kroemer 2004). More stringent definitions were used to further shortlist the identified FMCR into 17 deletions and 11 amplifications. Cancer related genes were enriched in the deleted (~22.5% of the total number of genes) and the amplified (~18.2% of the total number of genes) candidate FMCR. In total 6 novel candidate CRC TSG (*GMCL1*, *FOLH1*, *CDK2AP1*, *NFKBIA*, *IFNAR1* and *IFNAR2*) and 5 novel candidate CRC oncogenes (*PRDM16*, *PAX7*, *TNS1*, *RPA3*, *KCNMA1*) were proposed. Survival analysis suggested that high amplitude amplifications of *KCNMA1* may be associated with poor survival. None of the other candidate FMCR/genes was significantly associated with patient survival. *NFKBIA* deletions were previously associated with poor survival of GBM patients but this was not observed in our cohort. Out of the 45 samples included in the common aberration analysis, survival data was only available for 33, which limits the power to detect any statistically significant association.

The microRNA miR211 was recently reported to target *KCNMA1*. Underexpression of miR211 was shown to have oncogenic effects in melanoma, through abnormal activation of *KCNMA1* (Mazar *et al.* 2010). A statistically significant inverse correlation was observed between *MIR211* deletions and *KCNMA1* amplification in our samples. Similarly, miR301a was recently reported to activate NF-κB in pancreatic cancer (Lu *et al.* 2011). A statistically significant inverse correlation was observed between miR301a amplifications and *NFKBIA* deletions in our samples. These results support the notion that the above mutually exclusive defects function in the same oncogenic pathways.

NFKBIA, one of the identified candidate TSG, was selected for further analysis in a proof of concept experiment. NFKBIA protein was shown to be relatively under-expressed in 3 of the 6 CRC cell lines investigated (SW480, CACO2 and COLO205). *NFKBIA*-siRNA experiments were performed to deplete NFKBIA protein in CRC cell lines HCT116, SW620 and HT29 in which NFKBIA is relatively expressed at normal to high levels. This was achieved significantly in HCT116 and SW620, however, the effect appeared to be of relatively short duration as the NFKBIA levels were lower 24 hr after transfection compared to 48 hr. Preliminary results from growth and colony formation experiments indicated a survival advantage of NFKBIA underexpression in the investigated cell lines which was consistent with what was recently shown in GBM cell lines (Bredel *et al.* 2010).

7.4 Study limitations

The CRC molecular classification experiment was limited by a relatively small sample size and the absence of MSI-H cases. The main limiting factor is the high cost associated with the detailed examination of the molecular characteristics, especially when using high throughput technologies such as array CGH. Other molecular classification studies that included high throughput technologies as part of their analysis were also confined to a similar sample size of 60-70 cases (Cheng *et al.* 2008, Derks *et al.* 2008). Multi-centre collaborative studies with large CRC cohorts will thus be essential to achieve more comprehensive molecular classification results. The Cancer Genome Atlas project (TCGA) led by NCI and the National Human Genome Research Institute (NHGRI) is performing a comprehensive genomic characterisation of ~20-25 cancer types including CRC. TCGA aims to characterise ~500 cases for each cancer type. The characterisation involves genome wide copy number, methylation, gene expression and microRNA profiling, in addition to whole exon sequencing (http://cancergenome.nih.gov/, accessed September, 2011). The results of this project should provide, in the near future, much more detailed genomic information for CRC and other cancer types.

Another main limitation of this project was the lack of CRC tumour and normal RNA samples. Messenger RNA would have been very useful, especially in investigating the expression levels of the proposed candidate driver genes. Although a strong association was previously established between copy number aberrations and mRNA expression levels, it is not always the case (Tsafrir *et al.* 2006, Sheffer *et al.* 2009, Brosens *et al.* 2010). If mRNA was available, a direct comparison of copy number aberrations and expression levels could have been undertaken for the candidate driver genes. Moreover, investigating mRNA expression levels is also essential as part of the molecular classification, especially given the recently proposed transcriptome instability in CRC (Sveen *et al.* 2011). MicroRNA samples would have been also useful in the project for several reasons, including confirmation of the expression levels of miR211 and miR301a in relation to their copy number aberrations and *KCNMA1* and *NFKBIA* levels respectively.

7.5 Future work

Thorough analysis of *CASP8* tag SNPs in CRC is in progress which may suggest further SNPs for follow up studies. Analysis of the second generation sequencing data encompassing *CASP8* CNV23598 in the 47 breast cancer samples might reveal the exact sequence of the area flanking CNV23598 which can allow more accurate genotyping and further investigation of a potential role of CNV23598 in CRC and breast cancer risk.

Further investigation and validation of the chromosomally unstable CIMP-H/MSS pathway is required. A larger sample set would allow confirmation of this molecular subtype. It would also allow us to examine any clinical associations of this subtype.

Recent studies have shown that cancer related microRNA genes are targeted by chromosomal defects in several cancer types including CRC (Zhang *et al.* 2006, Diosdado *et al.* 2009). Further analysis of the aCGH data is required to investigate the presence of microRNA genes within the identified FMCR which can potentially be used to identify novel cancer related microRNAs.

Preliminary functional studies results have supported the predicted role of *NFKBIA* as a TSG in CRC. However, these studies should be replicated. The siRNA experiments suffer from having only a short-term effect on NFKBIA protein expression. This could affect the scale and reproducibility of the results. An alternative approach to investigate the role of *NFKBIA* in CRC would be to overexpress NFKBIA in cell lines such as, CACO2, SW480 and COLO205, where NFKBIA was shown to be underexpressed. Stable transfection and overexpression of NFKBIA in these cell lines may provide a better model for determining the role of *NFKBIA* as a TSG in CRC. This model will also be more suitable to conduct more experiments and investigate the effect of NFKBIA on resistance to chemotherapeutic agents. It is also important to examine *NFKBIA* copy number in the same cell lines in order to elucidate whether heterozygous deletions of the gene were also responsible for lower expression levels.

NF-κB is known to be constitutively activated in several cancer types including CRC, however, the mechanisms behind this are mainly unknown (Kojima *et al.* 2004, Lu *et al.* 2011, Rinkenbaugh and Baldwin 2011). *NFKBIA* deletions were predicted to promote tumourigenesis in GBM through the aberrant activation of NF-κB (Bredel *et al.* 2010), however, NF-κB activation was not investigated. For potential clinical and therapeutic studies, it is necessary to understand the mode of action of *NFKBIA* deletions and to confirm its effect on NF-κB activation. In lung cancer cell lines *NFKBIA*-siRNA were shown to affect p-RELA (activated NF-κB subunit) levels (Bivona *et al.* 2011). However, we could not establish this in our tested CRC cell lines, because the levels of p-RELA were very low using Western blot. Considering that NF-κB was previously shown to be constitutively active in CRC (Kojima *et al.* 2004). There might be a technical reason for not detecting p-RELA. When activated, RELA is translocated to the nucleus and phosphorylated. Our protein extraction procedure involved passing the protein extracts through a fine-gauge needle ~30 times in order to release nuclear proteins. It is possible that this procedure was not sufficient. Another approach will be to try including a sonication step in the protein extraction protocol to try to release nuclear proteins more efficiently. Immunohistochemical staining of p-RELA was previously shown to be effective in determining NF-κB activity in CRC samples (Kojima *et al.* 2004). Therefore, this method could also be used to investigate NF-κB levels in the CRC cell lines.

Functional studies (similar to the *NFKBIA* experiments) will be required to validate the other proposed candidate driver genes. Protein expression levels of the candidate CRC driver genes could be investigated in a panel of CRC cell lines, and based on the results of these experiments, further functional studies can be performed to investigate the role of the promising genes in CRC development. The possible association between *KCNMA1* and CRC prognosis should be examined in a large CRC cohort. This could be achieved relatively cheaply and quickly using a panel of TaqMan probes for the *KCNMA1* gene. Validating this association is important to propose *KCNMA1* as a prognostic marker and as a possible therapeutic target in CRC.

BIBLIOGRAPHY

Abubaker, J., Bavi, P., Al-Harbi, S., *et al.* (2008). Clinicopathological analysis of colorectal cancers with PIK3CA mutations in Middle Eastern population. *Oncogene,* 27**,** 3539-45.

Ahuja, N., Mohan, A. L., Li, Q., *et al.* (1997). Association between CpG island methylation and microsatellite instability in colorectal cancer. *Cancer research,* 57**,** 3370-4.

Allegra, C. J., Jessup, J. M., Somerfield, M. R., *et al.* (2009). American Society of Clinical Oncology provisional clinical opinion: testing for KRAS gene mutations in patients with metastatic colorectal carcinoma to predict response to anti-epidermal growth factor receptor monoclonal antibody therapy. *Journal of clinical oncology : official journal of the American Society of Clinical Oncology,* 27**,** 2091-6.

Alnemri, E. S., Livingston, D. J., Nicholson, D. W., *et al.* (1996). Human ICE/CED-3 protease nomenclature. *Cell,* 87**,** 171.

Amundadottir, L. T., Sulem, P., Gudmundsson, J., *et al.* (2006). A common variant associated with prostate cancer in European and African populations. *Nature genetics,* 38**,** 652-8.

Andersen, C. L., Lamy, P., Thorsen, K., *et al.* (2010). Frequent genomic loss at chr16p13.2 is associated with poor prognosis in colorectal cancer. *Int J Cancer*.

Arndt, G. M., Dossey, L., Cullen, L. M., *et al.* (2009). Characterization of global microRNA expression reveals oncogenic potential of miR-145 in metastatic colorectal cancer. *BMC Cancer,* 9**,** 374.

Baba, Y., Nosho, K., Shima, K., *et al.* (2009). Aurora-A expression is independently associated with chromosomal instability in colorectal cancer. *Neoplasia,* 11**,** 418-25.

Bacher, J. W., Flanagan, L. A., Smalley, R. L., *et al.* (2004). Development of a fluorescent multiplex assay for detection of MSI-High tumors. *Dis Markers,* 20**,** 237-50.

Balaguer, F., Link, A., Lozano, J. J., *et al.* (2010). Epigenetic silencing of miR-137 is an early event in colorectal carcinogenesis. *Cancer research,* 70**,** 6609-18.

Balmain, A., Gray, J. & Ponder, B. (2003). The genetics and genomics of cancer. *Nature genetics,* 33 Suppl**,** 238-44.

Barault, L., Charon-Barra, C., Jooste, V., *et al.* (2008). Hypermethylator phenotype in sporadic colon cancer: study on a population-based series of 582 cases. *Cancer research,* 68**,** 8541-6.

Barr, F. G., Nauta, L. E., Davis, R. J., *et al.* (1996). In vivo amplification of the PAX3-FKHR and PAX7-FKHR fusion genes in alveolar rhabdomyosarcoma. *Hum Mol Genet,* 5**,** 15-21.

Barrett, J. H., Smith, G., Waxman, R., *et al.* (2003). Investigation of interaction between N-acetyltransferase 2 and heterocyclic amines as potential risk factors for colorectal cancer. *Carcinogenesis,* 24**,** 275-82.

Basseres, D. S., Ebbs, A., Levantini, E., *et al.* (2010). Requirement of the NF-kappaB subunit p65/RelA for K-Ras-induced lung tumorigenesis. *Cancer research,* 70**,** 3537-46.

Beckmann, J. S., Estivill, X. & Antonarakis, S. E. (2007). Copy number variants and genetic traits: closer to the resolution of phenotypic to genotypic variability. *Nat Rev Genet,* 8**,** 639-46.

Bedi, A., Pasricha, P. J., Akhtar, A. J., *et al.* (1995). Inhibition of Apoptosis during Development of Colorectal Cancer. *Cancer Research,* 55**,** 1811-1816.

Ben-Dor, A., Lipson, D., Tsalenko, A., *et al.* (2007). Framework for identifying common aberrations in DNA copy number data. *Proceedings of the 11th* *annual international conference on Research in computational molecular biology,* 4453**,** 122-136.

Beroud, C. & Soussi, T. (1996). APC gene: database of germline and somatic mutations in human tumors and cell lines. *Nucleic acids research,* 24**,** 121-4.

Beroukhim, R., Mermel, C. H., Porter, D.*, et al.* (2010). The landscape of somatic copy-number alteration across human cancers. *Nature,* 463**,** 899-905.

Bertagnolli, M. M., Niedzwiecki, D., Compton, C. C.*, et al.* (2009). Microsatellite instability predicts improved response to adjuvant therapy with irinotecan, fluorouracil, and leucovorin in stage III colon cancer: Cancer and Leukemia Group B Protocol 89803. *Journal of clinical oncology : official journal of the American Society of Clinical Oncology,* 27**,** 1814-21.

Bertoni, F., Codegoni, A. M., Furlan, D.*, et al.* (1999). CHK1 frameshift mutations in genetically unstable colorectal and endometrial cancers. *Genes, chromosomes & cancer,* 26**,** 176-80.

Bivona, T. G., Hieronymus, H., Parker, J.*, et al.* (2011). FAS and NF-kappaB signalling modulate dependence of lung cancers on mutant EGFR. *Nature,* 471**,** 523-6.

Bloch, M., Ousingsawat, J., Simon, R.*, et al.* (2007). KCNMA1 gene amplification promotes tumor cell proliferation in human prostate cancer. *Oncogene,* 26**,** 2525-34.

Bohanes, P., Labonte, M. J., Winder, T.*, et al.* (2011). Predictive molecular classifiers in colorectal cancer. *Seminars in oncology,* 38**,** 576-87.

Boland, C. R., Thibodeau, S. N., Hamilton, S. R.*, et al.* (1998). A National Cancer Institute Workshop on Microsatellite Instability for Cancer Detection and Familial Predisposition: Development of International Criteria for the Determination of Microsatellite Instability in Colorectal Cancer. *Cancer Research,* 58**,** 5248-5257.

Bonfield, J. K., Rada, C. & Staden, R. (1998). Automated detection of point mutations using fluorescent sequence trace subtraction. *Nucleic Acids Res,* 26**,** 3404-9.

Bonizzi, G. & Karin, M. (2004). The two NF-kappaB activation pathways and their role in innate and adaptive immunity. *Trends in immunology,* 25**,** 280-8.

Bourdon, J.-C. (2007). p53 and its isoforms in cancer. *British Journal of Cancer,* 97**,** 277-282.

Brakensiek, K., Wingen, L. U., Langer, F.*, et al.* (2007). Quantitative high-resolution CpG island mapping with Pyrosequencing reveals disease-specific methylation patterns of the CDKN2B gene in myelodysplastic syndrome and myeloid leukemia. *Clinical chemistry,* 53**,** 17-23.

Bredel, M., Scholtens, D. M., Yadav, A. K.*, et al.* (2010). NFKBIA Deletion in Glioblastomas. *N Engl J Med*.

Brink, M., De Goeij, A. F., Weijenberg, M. P.*, et al.* (2003). K-ras oncogene mutations in sporadic colorectal cancer in The Netherlands Cohort Study. *Carcinogenesis,* 24**,** 703-10.

Broderick, P., Carvajal-Carmona, L., Pittman, A. M.*, et al.* (2007). A genome-wide association study shows that common alleles of SMAD7 influence colorectal cancer risk. *Nature Genetics,* 39**,** 1315-1317.

Brosens, R. P., Haan, J. C., Carvalho, B.*, et al.* (2010). Candidate driver genes in focal chromosomal aberrations of stage II colon cancer. *J Pathol,* 221**,** 411-24.

Brueck, C., Song, S. & Collins, J. (2007). Oligonucleotide Array CGH Analysis of a Robust Whole Genome Amplification Method. *BioTechniques,* 42**,** 230-233.

Butler, L. M., Hewett, P. J., Fitridge, R. A.*, et al.* (1999). Deregulation of Apoptosis in Colorectal Carcinoma: Theoretical and Therapeutic Implications. *Australian and New Zealand Journal of Surgery,* 69**,** 88-94.

Cabannes, E., Khan, G., Aillet, F.*, et al.* (1999). Mutations in the IkBa gene in Hodgkin's disease suggest a tumour suppressor role for IkappaBalpha. *Oncogene,* 18**,** 3063-70.

Camp, E. R., Li, J., Minnich, D. J.*, et al.* (2004). Inducible nuclear factor-kappaB activation contributes to chemotherapy resistance in gastric cancer. *Journal of the American College of Surgeons,* 199**,** 249-58.

Carvalho, B., Postma, C., Mongera, S.*, et al.* (2009). Multiple putative oncogenes at the chromosome 20q amplicon contribute to colorectal adenoma to carcinoma progression. *Gut,* 58**,** 79-89.

Chang, S.-C., Lin, J.-K., Yang, S. H.*, et al.* (2006). Relationship between genetic alterations and prognosis in sporadic colorectal cancer. *International Journal of Cancer,* 118**,** 1721–1727.

Chen, H., Duncan, I. C., Bozorgchami, H.*, et al.* (2002). Tensin1 and a previously undocumented family member, tensin2, positively regulate cell migration. *Proceedings of the National Academy of Sciences of the United States of America,* 99**,** 733-8.

Chen, H. I., Hsu, F. H., Jiang, Y.*, et al.* (2008). A probe-density-based analysis method for array CGH data: simulation, normalization and centralization. *Bioinformatics,* 24**,** 1749-56.

Cheng, S., Fockler, C., Barnes, W. M.*, et al.* (1994). Effective amplification of long targets from cloned inserts and human genomic DNA. *Proceedings of the National Academy of Sciences of the United States of America,* 91**,** 5695-9.

Cheng, Y.-W., Pincas, H., Bacolod, M. D.*, et al.* (2008). CpG Island Methylator Phenotype Associates with Low-Degree Chromosomal Abnormalities in Colorectal Cancer. *Cancer Research,* 14**,** 6005-6013.

Choong, M. K. & Tsafnat, G. (2011). Genetic and Epigenetic Biomarkers of Colorectal Cancer. *Clinical gastroenterology and hepatology : the official clinical practice journal of the American Gastroenterological Association*.

Chung, C. C. & Chanock, S. J. (2011). Current status of genome-wide association studies in cancer. *Human genetics,* 130**,** 59-78.

Compton, C. C. & Greene, F. L. (2004). The staging of colorectal cancer: 2004 and beyond. *CA: a cancer journal for clinicians,* 54**,** 295-308.

Cox, A., Dunning, A. M., Garcia-Closas, M.*, et al.* (2007). A common coding variant in CASP8 is associated with breast cancer risk. *Nature Genetics,* 39**,** 352 - 358.

Crawford, G. E., Holt, I. E., Whittle, J.*, et al.* (2006). Genome-wide mapping of DNase hypersensitive sites using massively parallel signature sequencing (MPSS). *Genome research,* 16**,** 123-31.

Cummins, J. M., He, Y., Leary, R. J.*, et al.* (2006). The colorectal microRNAome. *Proceedings of the National Academy of Sciences of the United States of America,* 103**,** 3687-92.

Curtin, K., Lin, W. Y., George, R.*, et al.* (2009a). Meta association of colorectal cancer confirms risk alleles at 8q24 and 18q21. *Cancer epidemiology, biomarkers & prevention : a publication of the American Association for Cancer Research, cosponsored by the American Society of Preventive Oncology,* 18**,** 616-21.

Curtin, K., Lin, W. Y., George, R.*, et al.* (2009b). Genetic variants in XRCC2: new insights into colorectal cancer tumorigenesis. *Cancer epidemiology, biomarkers & prevention : a publication of the American Association for Cancer Research, cosponsored by the American Society of Preventive Oncology,* 18**,** 2476-84.

Cybulski, C., Gorski, B., Huzarski, T.*, et al.* (2004). CHEK2 is a multiorgan cancer susceptibility gene. *American journal of human genetics,* 75**,** 1131-5.

Davies, H., Bignell, G. R., Cox, C.*, et al.* (2002). Mutations of the BRAF gene in Human cancer. *Nature,* 417**,** 949-954.

De La Chapelle, A. (2004). Genetic Predisposition to colorectal cancer. *Nature Reviews Cancer,* 4**,** 769-780.

Derks, S., Postma, C., Carvalho, B.*, et al.* (2008). Integrated analysis of chromosomal, microsatellite and epigenetic instability in colorectal cancer identifies specific associations between promoter methylation of pivotal tumour suppressor and DNA repair genes and specific chromosomal alterations. *Carcinogenesis,* 29**,** 434-439.

Deschoolmeester, V., Baay, M., Specenier, P.*, et al.* (2010). A review of the most promising biomarkers in colorectal cancer: one step closer to targeted therapy. *The oncologist,* 15**,** 699-731.

Devos, L., Chanson, A., Liu, Z.*, et al.* (2008). Associations between single nucleotide polymorphisms in folate uptake and metabolizing genes with blood folate, homocysteine, and DNA uracil concentrations. *The American journal of clinical nutrition,* 88**,** 1149-58.

Di Fiore, F., Sesboue, R., Michel, P.*, et al.* (2010). Molecular determinants of anti-EGFR sensitivity and resistance in metastatic colorectal cancer. *British journal of cancer,* 103**,** 1765-72.

Di Nicolantonio, F., Martini, M., Molinari, F.*, et al.* (2008). Wild-type BRAF is required for response to panitumumab or cetuximab in metastatic colorectal cancer. *Journal of clinical oncology : official journal of the American Society of Clinical Oncology,* 26**,** 5705-12.

Diep, C. B., Kleivi, K., Ribeiro, F. R.*, et al.* (2006). The order of genetic events associated with colorectal cancer progression inferred from meta-analysis of copy number changes. *Genes, chromosomes & cancer,* 45**,** 31-41.

Diosdado, B., Van De Wiel, M. A., Terhaar Sive Droste, J. S.*, et al.* (2009). MiR-17-92 cluster is associated with 13q gain and c-myc expression during colorectal adenoma to adenocarcinoma progression. *Br J Cancer,* 101**,** 707-14.

Dong, C. & Hemminki, K. (2001). Modification of cancer risks in offspring by sibling and parental cancers from 2,112,616 nuclear families. *International journal of cancer. Journal international du cancer,* 92**,** 144-50.

Dyrso, T., Li, J., Wang, K.*, et al.* (2011). Identification of chromosome aberrations in sporadic microsatellite stable and unstable colorectal cancers using array comparative genomic hybridization. *Cancer genetics,* 204**,** 84-95.

Earnshaw, W. C., Martins, L. M. & Kaufmann, S. H. (1999). Mammalian caspases: structure, activation, substrates, and functions during apoptosis. *Annual review of biochemistry,* 68**,** 383-424.

Easton, D. F., Pooley, K. A., Dunning, A. M.*, et al.* (2007). Genome-wide association study identifies novel breast cancer susceptibility loci. *Nature,* 447**,** 1087-93.

Economopoulos, K. P. & Sergentanis, T. N. (2010). GSTM1, GSTT1, GSTP1, GSTA1 and colorectal cancer risk: a comprehensive meta-analysis. *European journal of cancer,* 46**,** 1617-31.

Ehrich, M., Zoll, S., Sur, S.*, et al.* (2007). A new method for accurate assessment of DNA quality after bisulfite treatment. *Nucleic acids research,* 35**,** e29.

Ellis, R. E., Yuan, J. Y. & Horvitz, H. R. (1991). Mechanisms and functions of cell death. *Annual review of cell biology,* 7**,** 663-98.

Elmore, S. (2007). Apoptosis: a review of programmed cell death. *Toxicologic pathology,* 35**,** 495-516.

Eshleman, J. R., Casey, G., Kochera, M. E.*, et al.* (1998). Chromosome number and structure both are markedly stable in RER colorectal cancers and are not destabilized by mutation of p53. *Oncogene,* 17**,** 719-725.

Fadeel, B. & Orrenius, S. (2005). Apoptosis: a basic biological phenomenon with wide-ranging implications in human disease. *Journal of Internal Medicine,* 258**,** 479-517.

Fearon, E. R. & Vogelstein, B. (1990). A genetic model for colorectal tumorigenesis. *Cell,* 61**,** 759-67.

Frazier, M. L., Xi, L., Zong, J.*, et al.* (2003). Association of the CpG island methylator phenotype with family history of cancer in patients with colorectal cancer. *Cancer research,* 63**,** 4805-8.

Gao, J., Pfeifer, D., He, L. J.*, et al.* (2007). Association of NFKBIA polymorphism with colorectal cancer risk and prognosis in Swedish and Chinese populations. *Scand J Gastroenterol,* 42**,** 345-50.

Geelen, C. M. M. V., Vries, E. G. E. D. & Jong, S. D. (2004). Lessons from TRAIL-resistance mechanisms in colorectal cancer cells: paving the road to patient-tailored therapy. *Drug Resistance Updates,* 7**,** 345–358.

Giovannucci, E., Rimm, E. B., Ascherio, A.*, et al.* (1995). Alcohol, low-methionine--low-folate diets, and risk of colon cancer in men. *Journal of the National Cancer Institute,* 87**,** 265-73.

Goel, A., Nagasaka, T., Arnold, C. N.*, et al.* (2007). The CpG island methylator phenotype and chromosomal instability are inversely correlated in sporadic colorectal cancer. *Gastroenterology,* 132**,** 127-38.

Goel, A. & Shin, S. K. (2008). CpG Island methylator phenotype in colorectal cancer: A current perspective. *Current Colorectal Cancer Reports,* 4**,** 77-83.

Goldgar, D. E., Easton, D. F., Cannon-Albright, L. A.*, et al.* (1994). Systematic population-based assessment of cancer risk in first-degree relatives of cancer probands. *Journal of the National Cancer Institute,* 86**,** 1600-8.

Greenman, C., Stephens, P., Smith, R.*, et al.* (2007). Patterns of somatic mutation in human cancer genomes. *Nature,* 446**,** 153-8.

Grenet, J., Teitz, T., Wei, T.*, et al.* (1999). Structure and chromosome localization of the human CASP8 gene. *Gene,* 226**,** 225-32.

Groden, J., Thliveris, A., Samowitz, W.*, et al.* (1991). Identification and characterization of the familial adenomatous polyposis coli gene. *Cell,* 66**,** 589-600.

Guastadisegni, C., Colafranceschi, M., Ottini, L.*, et al.* (2010). Microsatellite instability as a marker of prognosis and response to therapy: a meta-analysis of colorectal cancer survival data. *European journal of cancer,* 46**,** 2788-98.

Gunes, C., Lichtsteiner, S., Vasserot, A. P.*, et al.* (2000). Expression of the hTERT gene is regulated at the level of transcriptional initiation and repressed by Mad1. *Cancer research,* 60**,** 2116-21.

Haiman, C. A., Garcia, R. R., Kolonel, L. N.*, et al.* (2008). A promoter polymorphism in the CASP8 gene is not associated with cancer risk. *Nature Genetics,* 40**,** 295-260.

Haiman, C. A., Marchand, L. L., Yamamato, J.*, et al.* (2007). A common genetic risk factor for colorectal and prostate cancer. *Nature Genetics,* 39**,** 954-956.

Hanahan, D. & Weinberg, R. A. (2000). The Hallmarks of Cancer. *Cell,* 100**,** 57–70.

Hanahan, D. & Weinberg, R. A. (2011). Hallmarks of cancer: the next generation. *Cell,* 144**,** 646-74.

Harris, H. (2008). Concerning the origin of malignant tumours by Theodor Boveri. Translated and annotated by Henry Harris. Preface. *Journal of cell science,* 121 Suppl 1**,** v-vi.

Hawkins, N., Norrie, M., Cheong, K.*, et al.* (2002). CpG island methylation in sporadic colorectal cancers and its relationship to microsatellite instability. *Gastroenterology,* 122**,** 1376-87.

Hayat, M. J., Howlader, N., Reichman, M. E.*, et al.* (2007). Cancer statistics, trends, and multiple primary cancer analyses from the Surveillance, Epidemiology, and End Results (SEER) Program. *The oncologist,* 12**,** 20-37.

Heid, C. A., Stevens, J., Livak, K. J.*, et al.* (1996). Real time quantitative PCR. *Genome research,* 6**,** 986-94.

Hemminki, K., Forsti, A. & Lorenzo Bermejo, J. (2009a). Surveying germline genomic landscape of breast cancer. *Breast cancer research and treatment,* 113**,** 601-3.

Hemminki, K., Forsti, A. & Lorenzo Bermejo, J. (2009b). Surveying the genomic landscape of colorectal cancer. *The American journal of gastroenterology,* 104**,** 789-90.

Hermsen, M., Postma, C., Baak, J.*, et al.* (2002). Colorectal adenoma to carcinoma progression follows multiple pathways of chromosomal instability. *Gastroenterology,* 123**,** 1109–1119.

Hinoue, T., Weisenberger, D. J., Pan, F.*, et al.* (2009). Analysis of the association between CIMP and BRAF in colorectal cancer by DNA methylation profiling. *PLoS One,* 4**,** e8357.

Hopkins-Donaldson, S., Bodmer, J. L., Bourloud, K. B.*, et al.* (2000). Loss of caspase-8 expression in highly malignant human neuroblastoma cells correlates with resistance to tumor necrosis factor-related apoptosis-inducing ligand-induced apoptosis. *Cancer research,* 60**,** 4315-9.

Horton, J. K. & Tepper, J. E. (2005). Staging of colorectal cancer: past, present, and future. *Clinical colorectal cancer,* 4**,** 302-12.

Houlston, R. S., Cheadle, J., Dobbins, S. E.*, et al.* (2010). Meta-analysis of three genome-wide association studies identifies susceptibility loci for colorectal cancer at 1q41, 3q26.2, 12q13.13 and 20q13.33. *Nature genetics,* 42**,** 973-7.

Huang Da, W., Sherman, B. T. & Lempicki, R. A. (2009). Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nature protocols,* 4**,** 44-57.

Huang, Y., Han, S., Li, Y.*, et al.* (2007). Different roles of MTHFR C677T and A1298C polymorphisms in colorectal adenoma and colorectal cancer: a meta-analysis. *Journal of human genetics,* 52**,** 73-85.

Huerta, S., Goulet, E. J. & Livingston, E. H. (2006). Colon cancer and apoptosis. *The American Journal of Surgery,* 191**,** 517-526.

Iacopetta, B. (2003). TP53 mutation in colorectal cancer. *Human mutation,* 21**,** 271-6.

Iacopetta, B., Grieu, F. & Amanuel, B. (2010). Microsatellite instability in colorectal cancer. *Asia-Pacific journal of clinical oncology,* 6**,** 260-9.

Igney, F. H. & Krammer, P. H. (2002). Death and Anti-Death: Tumour Resistance to Apoptosis. *Nature Reviews Cancer,* 2**,** 277-288.

Ikenoue, T., Kanai, F., Hikiba, Y.*, et al.* (2005). Functional analysis of PIK3CA gene mutations in human colorectal cancer. *Cancer research,* 65**,** 4562-7.

Inoue, J., Gohda, J., Akiyama, T.*, et al.* (2007). NF-kappaB activation in development and progression of cancer. *Cancer science,* 98**,** 268-74.

Ionov, Y., Peinado, M. A., Malkhosyan, S.*, et al.* (1993). Ubiquitous somatic mutations in simple repeated sequences reveal a new mechanism for colonic carcinogenesis. *Nature,* 363**,** 558-61.

Ishimura, N. & Gores, G. J. (2005). The death receptor TRAIL in cancer cell apoptosis. *Annals of Cancer Research and Therapy,* 13**,** 1-10.

Issa, J.-P. (2008). Colon Cancer: It's CIN or CIMP. *Cancer Research,* 14**,** 5939-5940.

Izzo, J. G., Malhotra, U., Wu, T. T.*, et al.* (2006). Association of activated transcription factor nuclear factor kappab with chemoradiation resistance and poor outcome in esophageal carcinoma. *Journal of clinical oncology : official journal of the American Society of Clinical Oncology,* 24**,** 748-54.

Jacobson, M. D., Weil, M. & Raff, M. C. (1997). Programmed cell death in animal development. *Cell,* 88**,** 347-54.

Jaeger, E., Webb, E., Howarth, K.*, et al.* (2008). Common genetic variants at the CRAC1 (HMPS) locus on chromosome 15q13.3 influence colorectal cancer risk. *Nature Genetics,* 40**,** 26-28.

Jass, J. R. (2007). Classification of colorectal cancer based on correlation of clinical, morphological and molecular features. *Histopathology,* 50**,** 113–130.

Jass, J. R. & Morson, B. C. (1987). Reporting colorectal cancer. *Journal of clinical pathology,* 40**,** 1016-23.

Jemal, A., Bray, F., Center, M. M., *et al.* (2011). Global cancer statistics. *CA: a cancer journal for clinicians,* 61**,** 69-90.

Jeuken, J. W., Cornelissen, S. J., Vriezen, M., *et al.* (2007). MS-MLPA: an attractive alternative laboratory assay for robust, reliable, and semiquantitative detection of MGMT promoter hypermethylation in gliomas. *Laboratory investigation; a journal of technical methods and pathology,* 87**,** 1055-65.

Jeyasuria, P., Subedi, K., Suresh, A., *et al.* (2011). Elevated Levels of Uterine Anti-Apoptotic Signaling May Activate NFKB and Potentially Confer Resistance to Caspase 3-Mediated Apoptotic Cell Death During Pregnancy in Mice. *Biology of reproduction,* 85**,** 417-24.

Jhawer, M., Goel, S., Wilson, A. J., *et al.* (2008). PIK3CA mutation/PTEN expression status predicts response of colon cancer cells to the epidermal growth factor receptor inhibitor cetuximab. *Cancer research,* 68**,** 1953-61.

Jo, W. S. & Carethers, J. M. (2006). Chemotherapeutic implications in microsatellite unstable colorectal cancer. *Cancer biomarkers : section A of Disease markers,* 2**,** 51-60.

Johnstone, R. W., Ruefli, A. A. & Lowe, S. W. (2002). Apoptosis: A Link Review between Cancer Genetics and Chemotherapy. *Cell,* 108**,** 153-164.

Jong, M. M. D., Nolte, I. M., Meerman, G. J. T., *et al.* (2002). Low-penetrance Genes and Their Involvement in Colorectal Cancer Susceptibility. *Cancer Epidemiology, Biomarkers & Prevention,* 11**,** 1332-1352.

Jover, R., Nguyen, T. P., Perez-Carbonell, L., *et al.* (2011). 5-Fluorouracil adjuvant chemotherapy does not increase survival in patients with CpG island methylator phenotype colorectal cancer. *Gastroenterology,* 140**,** 1174-81.

Kabbarah, O., Nogueira, C., Feng, B., *et al.* (2010). Integrative genome comparison of primary and metastatic melanomas. *PLoS One,* 5**,** e10770.

Kambara, T., Simms, L. A., Whitehall, V. L., *et al.* (2004). BRAF mutation is associated with DNA methylation in serrated polyps and cancers of the colorectum. *Gut,* 53**,** 1137-44.

Karapetis, C. S., Khambata-Ford, S., Jonker, D. J., *et al.* (2008). K-ras mutations and benefit from cetuximab in advanced colorectal cancer. *The New England journal of medicine,* 359**,** 1757-65.

Karin, M., Cao, Y., Greten, F. R., *et al.* (2002). NF-kappaB in cancer: from innocent bystander to major culprit. *Nature reviews. Cancer,* 2**,** 301-10.

Karpinski, P., Myszka, A., Ramsey, D., *et al.* (2010). Polymorphisms in methyl-group metabolism genes and risk of sporadic colorectal cancer with relation to the CpG island methylator phenotype. *Cancer epidemiology,* 34**,** 338-44.

Kawasaki, H., Altieri, D. C., Lu, C.-D., *et al.* (1998). Inhibition of Apoptosis by Survivin Predicts Shorter Survival Rates in Colorectal Cancer. *Cancer Research,* 58**,** 5071-5074.

Kerr D J, Y. a. a. H. R. (ed.) 2001. *ABC of Colorectal Cancer*: BMJ Books.

Kerr, J. F., Wyllie, A. H. & Currie, A. R. (1972). Apoptosis: a basic biological phenomenon with wide-ranging implications in tissue kinetics. *British journal of cancer,* 26**,** 239-57.

Khaitan, D., Sankpal, U. T., Weksler, B., *et al.* (2009). Role of KCNMA1 gene in breast cancer invasion and metastasis to brain. *BMC Cancer,* 9**,** 258.

Kim, G. P., Colangelo, L. H., Wieand, H. S., *et al.* (2007). Prognostic and predictive roles of high-degree microsatellite instability in colon cancer: a National Cancer Institute-National Surgical Adjuvant Breast and Bowel Project Collaborative Study. *Journal of clinical oncology : official journal of the American Society of Clinical Oncology,* 25**,** 767-72.

Kim, H. S., Lee, J. W., Soung, Y. H., *et al.* (2003). Inactivating mutations of caspase-8 gene in colorectal carcinomas. *Gastroenterology,* 125**,** 708-715.

Kim, J. C., Choi, J. S., Roh, S. A.*, et al.* (2010). Promoter methylation of specific genes is associated with the phenotype and progression of colorectal adenocarcinomas. *Annals of surgical oncology,* 17**,** 1767-76.

Kinzler, K. W. & Vogelstein, B. (1996). Lessons from hereditary colorectal cancer. *Cell,* 87**,** 159-70.

Kohler, E. M., Derungs, A., Daum, G.*, et al.* (2008). Functional definition of the mutation cluster region of adenomatous polyposis coli in colorectal tumours. *Human molecular genetics,* 17**,** 1978-87.

Kojima, M., Morisaki, T., Sasaki, N.*, et al.* (2004). Increased nuclear factor-kB activation in human colorectal carcinoma and its correlation with tumor progression. *Anticancer research,* 24**,** 675-81.

Kozma, L., Kiss, I., Szakall, S.*, et al.* (1994). Investigation of c-myc oncogene amplification in colorectal cancer. *Cancer letters,* 81**,** 165-9.

Krappmann, D., Emmerich, F., Kordes, U.*, et al.* (1999). Molecular mechanisms of constitutive NF-kappaB/Rel activation in Hodgkin/Reed-Sternberg cells. *Oncogene,* 18**,** 943-53.

Kumar, K., Brim, H., Giardiello, F.*, et al.* (2009). Distinct BRAF (V600E) and KRAS mutations in high microsatellite instability sporadic colorectal cancer in African Americans. *Clinical cancer research : an official journal of the American Association for Cancer Research,* 15**,** 1155-61.

Lares, M. R., Rossi, J. J. & Ouellet, D. L. (2010). RNAi and small interfering RNAs in human disease therapeutic applications. *Trends in biotechnology,* 28**,** 570-9.

Lascorz, J., Forsti, A., Chen, B.*, et al.* (2010). Genome-wide association study for colorectal cancer identifies risk polymorphisms in German familial cases and implicates MAPK signalling pathways in disease susceptibility. *Carcinogenesis,* 31**,** 1612-9.

Lassmann, S., Weis, R., Makowiec, F.*, et al.* (2007). Array CGH identifies distinct DNA copy number profiles of oncogenes and tumor suppressor genes in chromosomal- and microsatellite-unstable sporadic colorectal carcinomas. *Journal of molecular medicine,* 85**,** 293-304.

Leary, R. J., Lin, J. C., Cummins, J.*, et al.* (2008). Integrated analysis of homozygous deletions, focal amplifications, and sequence alterations in breast and colorectal cancers. *Proc Natl Acad Sci U S A,* 105**,** 16224-9.

Lerman, C. & Shields, A. E. (2004). Genetic testing for cancer susceptibility: the promise and the pitfalls. *Nature Reviews Cancer,* 4**,** 235-241.

Lichtenstein, P., Holm, N. V., Verkasalo, P. K.*, et al.* (2000). Environmental and Heritable Factors in the Causation of Cancer *The New England Journal of Medicine,* 343**,** 78-85.

Liedtke, C., Groger, N., Manns, M. P.*, et al.* (2003). The human caspase-8 promoter sustains basal activity through SP1 and ETS-like transcription factors and can be up-regulated by a p53-dependent mechanism. *J Biol Chem,* 278**,** 27593-604.

Lind, D. S., Hochwald, S. N., Malaty, J.*, et al.* (2001). Nuclear factor-kappa B is upregulated in colorectal cancer. *Surgery,* 130**,** 363-9.

Liu, L., Chen, L., Xu, Y.*, et al.* (2010). microRNA-195 promotes apoptosis and suppresses tumorigenicity of human colorectal cancer cells. *Biochem Biophys Res Commun,* 400**,** 236-40.

Liu, M. & Chen, H. (2010). The role of microRNAs in colorectal cancer. *Journal of genetics and genomics = Yi chuan xue bao,* 37**,** 347-58.

Loeb, L. A. (1991). Mutator Phenotype May Be Required for Multistage Carcinogenesis. *Cancer Research,* 51**,** 3075-3079.

Loeb, L. A. & Harris, C. C. (2008). Advances in chemical carcinogenesis: a historical review and prospective. *Cancer research,* 68**,** 6863-72.

Longley, D., Wilson, T., Mcewan, M.*, et al.* (2006). c-FLIP inhibits chemotherapy-induced colorectal cancer cell death. *Oncogene,* 25**,** 838–848.

Lowe, S. W. & Lin, A. W. (2000). Apoptosis in cancer. *Carcinogenesis,* 21**,** 485-495.

Lu, Z., Li, Y., Takwi, A.*, et al.* (2011). miR-301a as an NF-kappaB activator in pancreatic cancer cells. *The EMBO journal,* 30**,** 57-67.

Luchtenborg, M., Weijenberg, M. P., Roemen, G. M.*, et al.* (2004). APC mutations in sporadic colorectal carcinomas from The Netherlands Cohort Study. *Carcinogenesis,* 25**,** 1219-26.

Luo, X., Budihardjo, I., Zou, H.*, et al.* (1998). Bid, a Bcl2 interacting protein, mediates cytochrome c release from mitochondria in response to activation of cell surface death receptors. *Cell,* 94**,** 481-90.

Lynch, H. T. & De La Chapelle, A. (1999). Genetic susceptibility to non-polyposis colorectal cancer. *Journal of medical genetics,* 36**,** 801-18.

Lynch, H. T. & De La Chapelle, A. (2003). Hereditary colorectal cancer. *The New England journal of medicine,* 348**,** 919-32.

Lynch, H. T. & Krush, A. J. (1971). Cancer family "G" revisited: 1895-1970. *Cancer,* 27**,** 1505-11.

Lynch, H. T., Shaw, M. W., Magnuson, C. W.*, et al.* (1966). Hereditary factors in cancer. Study of two large midwestern kindreds. *Archives of internal medicine,* 117**,** 206-12.

Ma, X., Becker Buscaglia, L. E., Barker, J. R.*, et al.* (2011). MicroRNAs in NF-kappaB signaling. *Journal of molecular cell biology,* 3**,** 159-66.

Macpherson, G., Healey, C. S., Teare, M. D.*, et al.* (2004). Association of a common variant of the CASP8 gene with reduced risk of breast cancer. *Journal of the National Cancer Institute,* 96**,** 1866-9.

Man, T. K., Lu, X. Y., Jaeweon, K.*, et al.* (2004). Genome-wide array comparative genomic hybridization analysis reveals distinct amplifications in osteosarcoma. *BMC Cancer,* 4**,** 45.

Marte, B. (2006). (1890) Cancer as a genetic disease. *Nature Milestones in Cancer*.

Martin, E. S., Tonon, G., Sinha, R.*, et al.* (2007). Common and distinct genomic events in sporadic colorectal cancer and diverse cancer types. *Cancer Res,* 67**,** 10736-43.

Masuhara, M., Nagao, K., Nishikawa, M.*, et al.* (2003). Enhanced degradation of MDM2 by a nuclear envelope component, mouse germ cell-less. *Biochemical and biophysical research communications,* 308**,** 927-32.

Mazar, J., Deyoung, K., Khaitan, D.*, et al.* (2010). The regulation of miRNA-211 expression and its role in melanoma cell invasiveness. *PLoS One,* 5**,** e13779.

Mccarroll, S. A. & Altshuler, D. M. (2007). Copy-number variation and association studies of human disease. *Nature genetics,* 39**,** S37-42.

Meijers-Heijboer, H., Wijnen, J., Vasen, H.*, et al.* (2003). The CHEK2 1100delC mutation identifies families with a hereditary breast and colorectal cancer phenotype. *American journal of human genetics,* 72**,** 1308-14.

Meylan, E., Dooley, A. L., Feldser, D. M.*, et al.* (2009). Requirement for NF-kappaB signalling in a mouse model of lung adenocarcinoma. *Nature,* 462**,** 104-7.

Migliore, L., Migheli, F., Spisni, R.*, et al.* (2011). Genetics, cytogenetics, and epigenetics of colorectal cancer. *Journal of biomedicine & biotechnology,* 2011**,** 792362.

Miki, Y., Swensen, J., Shattuck-Eidens, D.*, et al.* (1994). A strong candidate for the breast and ovarian cancer susceptibility gene BRCA1. *Science,* 266**,** 66-71.

Miyaki, M., Konishi, M., Kikuchi-Yanoshita, R.*, et al.* (1994). Characteristics of somatic mutation of the adenomatous polyposis coli gene in colorectal tumors. *Cancer Res,* 54**,** 3011-20.

Miyoshi, Y., Nagase, H., Ando, H.*, et al.* (1992). Somatic mutations of the APC gene in colorectal tumors: mutation cluster region in the APC gene. *Hum Mol Genet,* 1**,** 229-33.

Mosmann, T. (1983). Rapid colorimetric assay for cellular growth and survival: application to proliferation and cytotoxicity assays. *Journal of immunological methods,* 65**,** 55-63.

Mullis, K., Faloona, F., Scharf, S.*, et al.* (1986). Specific enzymatic amplification of DNA in vitro: the polymerase chain reaction. *Cold Spring Harbor symposia on quantitative biology,* 51 Pt 1**,** 263-73.

Murphy, K. M., Zhang, S., Geiger, T.*, et al.* (2006). Comparison of the microsatellite instability analysis system and the Bethesda panel for the determination of microsatellite instability in colorectal cancers. *The Journal of molecular diagnostics : JMD,* 8**,** 305-11.

Muzio, M., Stockwell, B. R., Stennicke, H. R.*, et al.* (1998). An induced proximity model for caspase-8 activation. *The Journal of biological chemistry,* 273**,** 2926-30.

Nagasaka, T., Sasamoto, H., Notohara, K.*, et al.* (2004). Colorectal cancer with mutation in BRAF, KRAS, and wild-type with respect to both oncogenes showing different patterns of DNA methylation. *J Clin Oncol,* 22**,** 4584-94.

Nishikori, M. (2005). Classical and Alternative NF-KB Activation Pathways and Their Roles in Lymphoid Malignancies. *Journal of clinical and experimental hematopathology,* 45**,** 15-24.

Nishizuka, S., Tamura, G., Terashima, M.*, et al.* (1998). Loss of heterozygosity during the development and progression of differentiated adenocarcinoma of the stomach. *The Journal of Pathology,* 185**,** 38–43.

Nosho, K., Kawasaki, T., Ohnishi, M.*, et al.* (2008). PIK3CA mutation in colorectal cancer: relationship with genetic and epigenetic alterations. *Neoplasia,* 10**,** 534-41.

Nussbaum, R. L., Mcinnes, R. R. & Willard, H. F. 2001. *In:* GRIGG, L. L. (ed.) *Genetics in Medicine.* New York: W.B. Saunders.

Nygren, A. O., Ameziane, N., Duarte, H. M.*, et al.* (2005). Methylation-specific MLPA (MS-MLPA): simultaneous detection of CpG methylation and copy number changes of up to 40 sequences. *Nucleic acids research,* 33**,** e128.

Ogino, S., Cantor, M., Kawasaki, T.*, et al.* (2006a). CpG island methylator phenotype (CIMP) of colorectal cancer is best characterised by quantitative DNA methylation analysis and prospective cohort studies. *Gut,* 55**,** 1000-6.

Ogino, S. & Goel, A. (2008). Molecular Classification and Correlates in Colorectal Cancer. *Journal of Molecular Diagnostics,* 10**,** 13-27.

Ogino, S., Kawasaki, T., Kirkner, G. J.*, et al.* (2007). Evaluation of Markers for CpG Island Methylator Phenotype (CIMP) in Colorectal Cancer by a Large Population-Based Sample. *Journal of Molecular Diagnostics,* 9**,** 305-314.

Ogino, S., Kawasaki, T., Kirkner, G. J.*, et al.* (2006b). CpG island methylator phenotype-low (CIMP-low) in colorectal cancer: possible associations with male sex and KRAS mutations. *The Journal of molecular diagnostics : JMD,* 8**,** 582-8.

Ogino, S., Nosho, K., Kirkner, G. J.*, et al.* (2009). CpG island methylator phenotype, microsatellite instability, BRAF mutation and clinical outcome in colon cancer. *Gut,* 58**,** 90-6.

Okada, H. & Mak, T. W. (2004). Pathways of apoptotic and non-apoptotic death in tumour cells. *Nature reviews. Cancer,* 4**,** 592-603.

Oliveira, C., Velho, S., Moutinho, C.*, et al.* (2007). KRAS and BRAF oncogenic mutations in MSS colorectal carcinoma progression. *Oncogene,* 26**,** 158-63.

Otsuka, T., Kohno, T., Mori, M.*, et al.* (1996). Deletion mapping of chromosome 2 in human lung carcinoma. *Genes, chromosomes & cancer,* 16**,** 113-9.

Ozakyol, A., Ozdemir, M. & Artan, S. (2006). Fish detected p53 deletion and N-MYC amplification in colorectal cancer. *Hepato-gastroenterology,* 53**,** 192-5.

Parkin, D. M., Bray, F., Ferlay, J.*, et al.* (2005). Global Cancer Statistics, 2002. *CA: A Cancer Journal for Clinicians,* 55**,** 74-108.

Pasche, B. & Yi, N. (2010). Candidate gene association studies: successes and failures. *Current opinion in genetics & development,* 20**,** 257-61.

Pearson, T. A. & Manolio, T. A. (2008). How to interpret a genome-wide association study. *JAMA : the journal of the American Medical Association,* 299**,** 1335-44.

Pikarsky, E., Porat, R. M., Stein, I.*, et al.* (2004). NF-kappaB functions as a tumour promoter in inflammation-associated cancer. *Nature,* 431**,** 461-6.

Pingoud-Meier, C., Lang, D., Janss, A. J.*, et al.* (2003a). Loss of caspase-8 protein expression correlates with unfavorable survival outcome in childhood medulloblastoma. *Clinical cancer research : an official journal of the American Association for Cancer Research,* 9**,** 6401-9.

Pingoud-Meier, C., Lang, D., Janss, A. J.*, et al.* (2003b). Loss of caspase-8 protein expression correlates with unfavorable survival outcome in childhood medulloblastoma. *Clin Cancer Res,* 9**,** 6401-9.

Pinkel, D. & Albertson, D. G. (2005). Array comparative genomic hybridization and its applications in cancer. *Nature genetics,* 37 Suppl**,** S11-7.

Pino, M. S. & Chung, D. C. (2010). The chromosomal instability pathway in colon cancer. *Gastroenterology,* 138**,** 2059-72.

Pittman, A. M., Broderick, P., Sullivan, K.*, et al.* (2008). CASP8 variants D302H and -652 6N ins/del do not influence the risk of colorectal cancer in the United Kingdom population. *Br J Cancer,* 98**,** 1434-6.

Popat, S. & Houlston, R. S. (2005). A systematic review and meta-analysis of the relationship between chromosome 18q genotype, DCC status and colorectal cancer prognosis. *European journal of cancer,* 41**,** 2060-70.

Popat, S., Hubner, R. & Houlston, R. S. (2005). Systematic review of microsatellite instability and colorectal cancer prognosis. *Journal of clinical oncology : official journal of the American Society of Clinical Oncology,* 23**,** 609-18.

Porcelli, L., Assaraf, Y. G., Azzariti, A.*, et al.* (2011). The Impact of Folate Status on the Efficacy of Colorectal Cancer Treatment. *Current drug metabolism*.

Poulogiannis, G., Ichimura, K., Hamoudi, R. A.*, et al.* (2010a). Prognostic relevance of DNA copy number changes in colorectal cancer. *The Journal of pathology,* 220**,** 338-47.

Poulogiannis, G., Mcintyre, R. E., Dimitriadi, M.*, et al.* (2010b). PARK2 deletions occur frequently in sporadic colorectal cancer and accelerate adenoma development in Apc mutant mice. *Proceedings of the National Academy of Sciences of the United States of America,* 107**,** 15145-50.

Pritchard, C. C. & Grady, W. M. (2011). Colorectal cancer molecular biology moves into clinical practice. *Gut,* 60**,** 116-29.

Pugh, T. J., Delaney, A. D., Farnoud, N.*, et al.* (2008). Impact of whole genome amplification on analysis of copy number variants. *Nucleic Acids Res,* 36**,** e80.

Rajagopalan, H., Nowak, M. A., Vogelstein, B.*, et al.* (2003). The significance of unstable chromosomes in colorectal cancer. *Nature reviews. Cancer,* 3**,** 695-701.

Rashid, A., Shen, L., Morris, J. S.*, et al.* (2001). CpG island methylation in colorectal adenomas. *The American journal of pathology,* 159**,** 1129-35.

Reed, J. C. (2000). Mechanisms of Apoptosis. *American Journal of Pathology,* 157**,** 1415-1430.

Ribic, C. M., Sargent, D. J., Moore, M. J.*, et al.* (2003). Tumor microsatellite-instability status as a predictor of benefit from fluorouracil-based adjuvant chemotherapy for colon cancer. *The New England journal of medicine,* 349**,** 247-57.

Rinkenbaugh, A. L. & Baldwin, A. S. (2011). Monoallelic deletion of NFKBIA in glioblastoma: when less is more. *Cancer Cell,* 19**,** 163-5.

Rodriguez, J. & Lazebnik, Y. (1999). Caspase-9 and APAF-1 form an active holoenzyme. *Genes & development,* 13**,** 3179-84.

Roth, A. D., Tejpar, S., Delorenzi, M.*, et al.* (2010). Prognostic role of KRAS and BRAF in stage II and III resected colon cancer: results of the translational study on the PETACC-3, EORTC 40993, SAKK 60-00 trial. *Journal of clinical oncology : official journal of the American Society of Clinical Oncology,* 28**,** 466-74.

Rozen, S. & Skaletsky, H. (2000). Primer3 on the WWW for general users and for biologist programmers. *Methods Mol Biol,* 132**,** 365-86.

Ryu, B.-K., Lee, M.-G., Chi, S.-G.*, et al.* (2001). Increased expression of cFLIPL in colonic adenocarcinoma. *Journal of Pathology,* 194**,** 15–19.

Sadahiro, S., Suzuki, T., Maeda, Y.*, et al.* (2010). Molecular determinants of folate levels after leucovorin administration in colorectal cancer. *Cancer chemotherapy and pharmacology,* 65**,** 735-42.

Samowitz, W. S., Albertsen, H., Herrick, J.*, et al.* (2005). Evaluation of a large, population-based sample supports a CpG island methylator phenotype in colon cancer. *Gastroenterology,* 129**,** 837-45.

Samowitz, W. S., Curtin, K., Ma, K. N.*, et al.* (2001a). Microsatellite instability in sporadic colon cancer is associated with an improved prognosis at the population level. *Cancer epidemiology, biomarkers & prevention : a publication of the American Association for Cancer Research, cosponsored by the American Society of Preventive Oncology,* 10**,** 917-23.

Samowitz, W. S., Holden, J. A., Curtin, K.*, et al.* (2001b). Inverse relationship between microsatellite instability and K-ras and p53 gene alterations in colon cancer. *The American journal of pathology,* 158**,** 1517-24.

Samuels, Y., Wang, Z., Bardelli, A.*, et al.* (2004). High frequency of mutations of the PIK3CA gene in human cancers. *Science,* 304**,** 554.

Sanger, F., Nicklen, S. & Coulson, A. R. (1977). DNA sequencing with chain-terminating inhibitors. *Proceedings of the National Academy of Sciences of the United States of America,* 74**,** 5463-7.

Sartore-Bianchi, A., Martini, M., Molinari, F.*, et al.* (2009). PIK3CA mutations in colorectal cancer are associated with clinical resistance to EGFR-targeted monoclonal antibodies. *Cancer research,* 69**,** 1851-7.

Scaffidi, C., Medema, J. P., Krammer, P. H.*, et al.* (1997). FLICE is predominantly expressed as two functionally active isoforms, caspase-8/a and caspase-8/b. *J Biol Chem,* 272**,** 26953-8.

Scartozzi, M., Bearzi, I., Pierantoni, C.*, et al.* (2007). Nuclear factor-kB tumor expression predicts response and survival in irinotecan-refractory metastatic colorectal cancer treated with cetuximab-irinotecan therapy. *Journal of clinical oncology : official journal of the American Society of Clinical Oncology,* 25**,** 3930-5.

Schee, K., Fodstad, O. & Flatmark, K. (2010). MicroRNAs as biomarkers in colorectal cancer. *The American journal of pathology,* 177**,** 1592-9.

Schouten, J. P., Mcelgunn, C. J., Waaijer, R.*, et al.* (2002). Relative quantification of 40 nucleic acid sequences by multiplex ligation-dependent probe amplification. *Nucleic acids research,* 30**,** e57.

Sethi, G., Ahn, K. S., Chaturvedi, M. M.*, et al.* (2007). Epidermal growth factor (EGF) activates nuclear factor-kappaB through IkappaBalpha kinase-independent but EGF receptor-kinase dependent tyrosine 42 phosphorylation of IkappaBalpha. *Oncogene,* 26**,** 7324-32.

Sheffer, M., Bacolod, M. D., Zuk, O.*, et al.* (2009). Association of survival and disease progression with chromosomal instability: A genomic exploration of colorectal cancer. *Proc Natl Acad Sci U S A,* 106**,** 7131-7136.

Shen, L., Toyota, M., Kondo, Y.*, et al.* (2007). Integrated genetic and epigenetic analysis identifies three different subclasses of colon cancer. *Proceedings of the National Academy of Sciences,* 104**,** 18654–18659.

Shephard, N. D., Abo, R., Rigas, S. H.*, et al.* (2009). A breast cancer risk haplotype in the caspase-8 gene. *Cancer research,* 69**,** 2724-8.

Shi, Y. (2004). Caspase activation: revisiting the induced proximity model. *Cell,* 117**,** 855-8.

Shimizu, S., Narita, M. & Tsujimoto, Y. (1999). Bcl-2 family proteins regulate the release of apoptogenic cytochrome c by the mitochondrial channel VDAC. *Nature,* 399**,** 483-7.

Shing, D. C., Trubia, M., Marchesi, F.*, et al.* (2007). Overexpression of sPRDM16 coupled with loss of p53 induces myeloid leukemias in mice. *The Journal of clinical investigation,* 117**,** 3696-707.

Shivapurkar, N., Toyooka, S., Eby, M. T.*, et al.* (2002a). Differential inactivation of caspase-8 in lung cancers. *Cancer biology & therapy,* 1**,** 65-9.

Shivapurkar, N., Toyooka, S., Eby, M. T.*, et al.* (2002b). Differential inactivation of caspase-8 in lung cancers. *Cancer Biol Ther,* 1**,** 65-9.

Sjoblom, T., Jones, S., Wood, L. D.*, et al.* (2006). The consensus coding sequences of human breast and colorectal cancers. *Science,* 314**,** 268-74.

Slattery, M. L., Lundgreen, A., Herrick, J. S.*, et al.* (2011). Variation in the CYP19A1 gene and risk of colon and rectal cancer. *Cancer Causes Control,* 22**,** 955-63.

Son, J.-W., Kang, H.-K., Chae, M. H.*, et al.* (2006). Polymorphisms in the caspase-8 gene and the risk of lung cancer. *Cancer Genetics and Cytogenetics,* 169**,** 121-127.

Soung, Y. H., Lee, J. W., Kim, S. Y.*, et al.* (2005). CASPASE-8 Gene Is Inactivated by Somatic Mutations in Gastric Carcinomas. *Cancer Research,* 65**,** 815-821.

Steel, M., Thompson, A. & Clayton, J. (1991). Genetic aspects of breast cancer. *British medical bulletin,* 47**,** 504-18.

Stratton, M. R., Campbell, P. J. & Futreal, P. A. (2009). The cancer genome. *Nature,* 458**,** 719-24.

Stupack, D. G., Teitz, T., Potter, M. D.*, et al.* (2006). Potentiation of neuroblastoma metastasis by loss of caspase-8. *Nature,* 439**,** 95-9.

Suh, J., Payvandi, F., Edelstein, L. C.*, et al.* (2002). Mechanisms of constitutive NF-kappaB activation in human prostate cancer cells. *The Prostate,* 52**,** 183-200.

Sun, T., Gao, Y., Tan, W.*, et al.* (2007). A six-nucleotide insertion-deletion polymorphism in the CASP8 promoter is associated with susceptibility to multiple cancers. *Nature Genetics,* 39**,** 605-613.

Suraweera, N., Duval, A., Reperant, M.*, et al.* (2002). Evaluation of tumor microsatellite instability using five quasimonomorphic mononucleotide repeats and pentaplex PCR. *Gastroenterology,* 123**,** 1804-11.

Sveen, A., Agesen, T. H., Nesbakken, A.*, et al.* (2011). Transcriptome instability in colorectal cancer identified by exon microarray analyses: Associations with splicing factor expression levels and patient survival. *Genome medicine,* 3**,** 32.

Takita, J., Yang, H. W., Chen, Y. Y.*, et al.* (2001). Allelic imbalance on chromosome 2q and alterations of the caspase 8 gene in neuroblastoma. *Oncogene,* 20**,** 4424-4432.

Talseth-Palmer, B. A., Bowden, N. A., Hill, A.*, et al.* (2008). Whole genome amplification and its impact on CGH array profiles. *BMC Res Notes,* 1**,** 56.

Tanaka, T., Watanabe, T., Kazama, Y.*, et al.* (2006). Chromosome 18q deletion and Smad4 protein inactivation correlate with liver metastasis: A study matched for T- and N- classification. *British journal of cancer,* 95**,** 1562-7.

Teitz, T., Wei, T., Valentine, M. B.*, et al.* (2000). Caspase 8 is deleted or silenced preferentially in childhood neuroblastomas with amplification of MYCN. *Nature Medicine,* 6 529-535.

Tenesa, A., Farrington, S. M., Prendergast, J. G.*, et al.* (2008). Genome-wide association scan identifies a colorectal cancer susceptibility locus on 11q23 and replicates risk loci at 8q24 and 18q21. *Nature genetics,* 40**,** 631-7.

Thibodeau, S. N., Bren, G. & Schaid, D. (1993). Microsatellite instability in cancer of the proximal colon. *Science,* 260**,** 816-9.

Thornberry, N. A., Rano, T. A., Peterson, E. P.*, et al.* (1997). A combinatorial approach defines specificities of members of the caspase family and granzyme B. Functional relationships established for key mediators of apoptosis. *The Journal of biological chemistry,* 272**,** 17907-11.

Toaldo, C., Pizzimenti, S., Cerbone, A.*, et al.* (2010). PPARgamma ligands inhibit telomerase activity and hTERT expression through modulation of the Myc/Mad/Max network in colon cancer cells. *J Cell Mol Med,* 14**,** 1347-57.

Tomlinson, I., Webb, E., Carvajal-Carmona, L.*, et al.* (2007). A genome-wide association scan of tag SNPs identifies a susceptibility variant for colorectal cancer at 8q24.21. *Nature Genetics,* 39**,** 984-988.

Tomlinson, I. P., Carvajal-Carmona, L. G., Dobbins, S. E.*, et al.* (2011). Multiple common susceptibility variants near BMP pathway loci GREM1, BMP4, and BMP2 explain part of the missing heritability of colorectal cancer. *PLoS genetics,* 7**,** e1002105.

Tomlinson, I. P., Webb, E., Carvajal-Carmona, L.*, et al.* (2008). A genome-wide association study identifies colorectal cancer susceptibility loci on chromosomes 10p14 and 8q23.3. *Nature Genetics,* 40**,** 623-630.

Toyota, M., Ahuja, N., Ohe-Toyota, M.*, et al.* (1999). CpG island methylator phenotype in colorectal cancer. *Proceedings of the National Academy of Sciences of the United States of America,* 96**,** 8681-6.

Toyota, M., Ohe-Toyota, M., Ahuja, N.*, et al.* (2000). Distinct genetic profiles in colorectal tumors with or without the CpG island methylator phenotype. *Proceedings of the National Academy of Sciences of the United States of America,* 97**,** 710-5.

Trojan, J., Brieger, A., Raedle, J.*, et al.* (2004). BAX and caspase-5 frameshift mutations and spontaneous apoptosis in colorectal cancer with microsatellite instability. *International journal of colorectal disease,* 19**,** 538-44.

Tsafrir, D., Bacolod, M., Selvanayagam, Z.*, et al.* (2006). Relationship of gene expression and chromosomal abnormalities in colorectal cancer. *Cancer research,* 66**,** 2129-37.

Umar, A., Boland, C. R., Terdiman, J. P.*, et al.* (2004). Revised Bethesda Guidelines for hereditary nonpolyposis colorectal cancer (Lynch syndrome) and microsatellite instability. *J Natl Cancer Inst,* 96**,** 261-8.

Vazquez, A., Bond, E. E., Levine, A. J.*, et al.* (2008). The genetics of the p53 pathway, apoptosis and cancer therapy. *Nature Reviews,* 7**,** 979-987.

Veeriah, S., Taylor, B. S., Meng, S.*, et al.* (2010). Somatic mutations of the Parkinson's disease-associated gene PARK2 in glioblastoma and other human malignancies. *Nature genetics,* 42**,** 77-82.

Velho, S., Oliveira, C., Ferreira, A.*, et al.* (2005). The prevalence of PIK3CA mutations in gastric and colon cancer. *Eur J Cancer,* 41**,** 1649-54.

Venkatachalam, R., Verwiel, E. T., Kamping, E. J.*, et al.* (2010). Identification of candidate predisposing copy number variants in familial and early-onset colorectal cancer patients. *Int J Cancer*.

Vilar, E. & Gruber, S. B. (2010). Microsatellite instability in colorectal cancer-the stable evidence. *Nature reviews. Clinical oncology,* 7**,** 153-62.

Vilar, E., Scaltriti, M., Balmana, J.*, et al.* (2008). Microsatellite instability due to hMLH1 deficiency is associated with increased cytotoxicity to irinotecan in human colorectal cancer cell lines. *British journal of cancer,* 99**,** 1607-12.

Vogelstein, B., Fearon, E. R., Hamilton, S. R.*, et al.* (1988). Genetic alterations during colorectal-tumor development. *The New England journal of medicine,* 319**,** 525-32.

Vogelstein, B. & Kinzler, K. W. (2004). Cancer genes and the pathways they control. *Nature medicine,* 10**,** 789-99.

Walther, A., Houlston, R. & Tomlinson, I. (2008). Association between chromosomal instability and prognosis in colorectal cancer: a meta-analysis. *Gut,* 57**,** 941-50.

Walther, A., Johnstone, E., Swanton, C.*, et al.* (2009). Genetic prognostic and predictive markers in colorectal cancer. *Nature Reviews Cancer,* 9**,** 489-499.

Wang, C. Y., Mayo, M. W. & Baldwin, A. S., Jr. (1996). TNF- and cancer therapy-induced apoptosis: potentiation by inhibition of NF-kappaB. *Science,* 274**,** 784-7.

Wang, H., Vardy, L. A., Tan, C. P.*, et al.* (2010). PCBP1 suppresses the translation of metastasis-associated PRL-3 phosphatase. *Cancer Cell,* 18**,** 52-62.

Wang, Z., Cummins, J. M., Shen, D.*, et al.* (2004). Three classes of genes mutated in colorectal cancers with chromosomal instability. *Cancer research,* 64**,** 2998-3001.

Weisenberger, D. J., Siegmund, K. D., Campan, M.*, et al.* (2006). CpG island methylator phenotype underlies sporadic microsatellite instability and is tightly associated with BRAF mutation in colorectal cancer. *Nature Genetics,* 38**,** 789-793.

Westra, J. L., Schaapveld, M., Hollema, H.*, et al.* (2005). Determination of TP53 mutation is more relevant than microsatellite instability status for the prediction of disease-free survival in adjuvant-treated stage III colon cancer patients. *Journal of clinical oncology : official journal of the American Society of Clinical Oncology,* 23**,** 5635-43.

Whiffin, N., Broderick, P., Lubbe, S. J.*, et al.* (2011). MLH1-93G>A is a risk factor for MSI colorectal cancer. *Carcinogenesis*.

Who 2008. The Global Burden of Disease: 2004 Update. Geneva: World Health Organisation.

Wong, K. M., Hudson, T. J. & Mcpherson, J. D. (2011). Unraveling the Genetics of Cancer: Genome Sequencing and Beyond. *Annual review of genomics and human genetics*.

Wood, L. D., Parsons, D. W., Jones, S.*, et al.* (2007). The genomic landscapes of human breast and colorectal cancers. *Science,* 318**,** 1108-13.

Woodford-Richens, K. L., Rowan, A. J., Gorman, P.*, et al.* (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 of the United States of America,* 98**,** 9719-23.

Wooster, R., Bignell, G., Lancaster, J.*, et al.* (1995). Identification of the breast cancer susceptibility gene BRCA2. *Nature,* 378**,** 789-92.

Yashiro, M., Hirakawa, K. & Boland, C. R. (2010). Mutations in TGFbeta-RII and BAX mediate tumor progression in the later stages of colorectal cancer with microsatellite instability. *BMC Cancer,* 10**,** 303.

Yin, J., Kong, D., Wang, S.*, et al.* (1997). Mutation of hMSH3 and hMSH6 mismatch repair genes in genetically unstable human colorectal and gastric carcinomas. *Human mutation,* 10**,** 474-8.

Yin, X. M., Wang, K., Gross, A.*, et al.* (1999). Bid-deficient mice are resistant to Fas-induced hepatocellular apoptosis. *Nature,* 400**,** 886-91.

Yuen, S. T., Davies, H., Chan, T. L.*, et al.* (2002). Similarity of the phenotypic patterns associated with BRAF and KRAS mutations in colorectal neoplasia. *Cancer Res,* 62**,** 6451-5.

Zanke, B. W., Greenwood, C. M., Rangrej, J.*, et al.* (2007). Genome-wide association scan identifies a colorectal cancer susceptibility locus on chromosome 8q24. *Nature genetics,* 39**,** 989-94.

Zhang, A., Wu, Y., Lai, H. W. L.*, et al.* (2004). Apoptosis – A Brief Review. *Neuroembryology,* 5**,** 47-59.

Zhang, K. X., Matsui, Y., Hadaschik, B. A.*, et al.* (2010). Down-regulation of type I interferon receptor sensitizes bladder cancer cells to vesicular stomatitis virus-induced cell death. *International journal of cancer. Journal international du cancer,* 127**,** 830-8.

Zhang, L., Huang, J., Yang, N.*, et al.* (2006). microRNAs exhibit high frequency genomic alterations in human cancer. *Proceedings of the National Academy of Sciences of the United States of America,* 103**,** 9136-41.

Zhao, Y., Sui, X. & Ren, H. (2010). From procaspase-8 to caspase-8: revisiting structural functions of caspase-8. *Journal of cellular physiology,* 225**,** 316-20.

Zhivotovsky, B. & Kroemer, G. (2004). Apoptosis and Genomic Instability. *Nature Reviews,* 5**,** 753-762.

Zolochevska, O. & Figueiredo, M. L. (2010a). Cell-cycle regulators cdk2ap1 and bicalutamide suppress malignant biological interactions between prostate cancer and bone cells. *Prostate*.

Zolochevska, O. & Figueiredo, M. L. (2010b). Novel tumor growth inhibition mechanism by cell cycle regulator cdk2ap1 involves antiangiogenesis modulation. *Microvasc Res,* 80**,** 324-31.

Zuzak, T. J., Steinhoff, D. F., Sutton, L. N.*, et al.* (2002). Loss of caspase-8 mRNA expression is common in childhood primitive neuroectodermal brain tumour/medulloblastoma. *European journal of cancer,* 38**,** 83-91.

APPENDICES

Appendix 1. Primers Sequences

Table 1 *CASP8* Primers

Primer	Primer Sequence 5'-3'	Length	GC %	Tm º C
CASP8-promoter Forward	TCAAGCTGTCTGCAGTCAGCAG	22	54.5	62.00
CASP8-promoter Reverse	CTCATACCAGTGTACAGGGAAC	22	50.0	60.25
CASP8-exon3F	CATCCCTCTTCTGAATGGTTGG	22	50.0	60.25
CASP8-exon3R	CAGAGAGAAACCACACAGGAG	21	52.4	59.80
CASP8-exon4F	GAAGCTTTAGAAGGCACTCTGC	22	50.0	60.25
CASP8-exon4R	AGAGGGACTGCTACCTCTTAC	21	52.4	59.80
CASP8-exon5F	AGAGGGAACTTGTCTGGTGTTC	22	50.0	60.25
CASP8-exon5R	ACCCAAGGCACTACTGTCTTTC	22	50.0	60.25
CASP8-exon6F	GGGGAGAACACTATTCAACCAG	22	50.0	60.25
CASP8-exon6R	AATGGTAAGTGGTCACCCTAGC	22	50.0	60.25
CASP8-exon7F	TAGACTTCTGTGTCACCACACC	22	50.0	60.25
CASP8-exon7R	CACATCTGTAACACTCGTGCTG	22	50.0	60.25
CASP8-exon8/9F	ATGGAATCGCTTCCCTAGTAGC	22	50.0	60.25
CASP8-exon8/9R	TGCATGTGGTAGAAGGCTGTG	21	52.4	59.80
CASP8-exon10F	TGCTTGCAGAATCTCTCTGGC	21	52.4	59.80
CASP8-exon10R	TCCAGTATCTTCATCTCGGGTC	22	50.0	60.25
CASP8-exon11F	GGTCCTGTGTGAAGGAAATAGG	22	50.0	60.25
CASP8-exon11R	TGCACTTAACAGCCCAACCTG	21	52.4	59.80
CASP8-exon12Fn	CAATGTTATGCCCACTGTGCTC	22	50.0	60.25
CASP8-exon12R	TGACCTGGGAAATGCAGCTATG	22	50.0	60.25
CASP8-exon13F	GCAGATGCGATGTCAATTCGAG	22	50.0	60.25
CASP8-exon13R	TGCTTCTCAGACCTTTGCCATG	22	50.0	60.25
Promoter-seqF1	TACAAGCTGCTGGCAGTCATG	21	52.4	59.80
Promoter-seqF2	TTATGAATGAGCCGAGGAAGGC	22	50.0	60.25
Promoter-seqF3	AACGACAACTCACAGTGCCAG	21	52.4	59.80
Promoter-seqF4	GCAGAGCTGGGATTTGAATCC	21	52.4	59.80
Exon13-seqF1	GTAGGTCTTGGTTTCGCACC	20	55.0	59.35
Exon13-seqF2	GTAGAGACAGGGTTTCACTGTG	22	50.0	60.25
Exon13-seqF3	GATTGCTTGAACCCAAGAGGTC	22	50.0	60.25
Promoter-seqR1	GGATTCAAATCCCAGCTCTGC	21	52.4	59.80
Promoter-seqR2	TTCCTGGCACTGTGAGTTGTC	21	52.4	59.80
Promoter-seqR3	TTAACGTCTCAGTGCCTTCCTC	22	50.0	60.25
Exon13-seqR1	AGTGCAGCGGTGTGAACATG	21	55.0	59.35
Exon13-seqR2	GCCTGTAATCCCAGCACTTTG	22	52.4	59.80
CASP8-exon10Fseq	AAGTGATCTGCCCATCTTGGC	22	52.4	59.80

Table 2 ***KRAS* Primers**

Primer	Primer Sequence 5'-3'	Length	GC%	Tm ° C
KRAS-exon1F	ACGATACACGTCTGCAGTCAAC	22	50.0	60.30
KRAS-exon1R	GCACAGAGAGTGAACATCATGG	22	50.0	60.30

Table 3 ***BRAF* Primers**

Primer	Primer Sequence 5'-3'	Length	GC%	Tm ° C
BRAF-exon15F	CCAGGAGTGCCAAGAGAATATC	22	50.0	60.30
BRAF-exon15R	AGTAACTCAGCAGCATCTCAGG	22	50.0	60.30

Table 4 ***TP53* Primers**

Primer	Sequence 5'-3'	Length	GC%	Tm ° C
TP53-exon4F	ACGTTCTGGTAAGGACAAGGG	21	52.4	59.8
TP53-exon4R	GACAGGAGTCAGAGATCACAC	21	52.4	59.8
TP53-exon5/6F	AAAGCTCCTGAGGTGTAGACG	21	52.4	59.8
TP53-exon5/6R	GGGAGGTCAAATAAGCAGCAG	21	52.4	59.8
TP53-exon7F	AAAAGGCCTCCCCTGCTTGC	20	60.0	61.4
TP53-exon7R	TGATGAGAGGTGGATGGGTAG	21	52.4	59.8
TP53-exon8/9F	AGCTTAGGCTCCAGAAAGGAC	21	52.4	59.8
TP53-exon8/9R	AGTTAGCTACAACCAGGAGCC	21	52.4	59.8
TP53-exon10F	GTCAGCTGTATAGGTACTTGAAG	22	43.5	58.9
TP53-exon10R	TGACCATGAAGGCAGGATGAG	21	52.4	59.8

Table 5 ***APC* Primers**

Primer	Primer Sequence 5'-3'	Length	GC%	Tm ° C
*APC*MCRf	GTTCTGCACAGAGTAGAAGTGG	22	50.0	60.3
*APC*MCRr	GTGATGACTTTGTTGGCATGGC	22	50.0	60.3
*APC*MCReqf	CTCCGTTCAGAGTGAACCATG	21	52.4	59.8
*APC*MCReqr	CATGGTTCACTCTGAACGGAG	21	52.4	59.8

Table 6 ***PIK3CA* Primers**

Primer	Primer Sequence 5'-3'	Length	GC%	Tm ° C
*PIK3CA*exon9F	CAGTTACTATTCTGTGACTGGTG	23	43.5	58.9
*PIK3CA*exon9R	TGCTGAGATCAGCCAAATTCAG	22	45.5	58.4
*PIK3CA*exon20F	TTGCTCCAAACTGACCAAACTG	22	45.5	58.4
*PIK3CA*exon20R	TGCAATTCCTATGCAATCGGTC	22	45.5	58.4

Table 7 CNV23598 primers

Primer	Primer Sequence 5'-3'	Length	GC%	Tm º C
Common forward	TAGAAGCCTGCAGAATCCAGC	21	52.4	59.80
Insertion reverse	GGATGGGCCATGATGACAATG	21	52.4	59.80
Deletion reverse	GAGATGTCAGCTCATAGATGGG	22	50.0	60.30
Fluorescent insertion reverse	CCTCCACTATTGTCCTGTGAC	21	52.4	59.80
Fluorescent deletion reverse	GAAGCTCTTCAAAGGTCGTGG	21	52.4	59.80
A/G SNP confirmation forward	GTCACAGGACAATAGTGGAGG	21	52.4	59.80
Forward deletion	TAGAAGCCTGCAGAATCCAGC	21	52.4	59.80
New insertion forward	AGCAGTACCGTCCAGCTTTG	20	55.0	59.35
Common reverse	GAGATGTCAGCTCATAGATGGG	22	50.0	60.30

Table 8 Pyrosequencing Primers

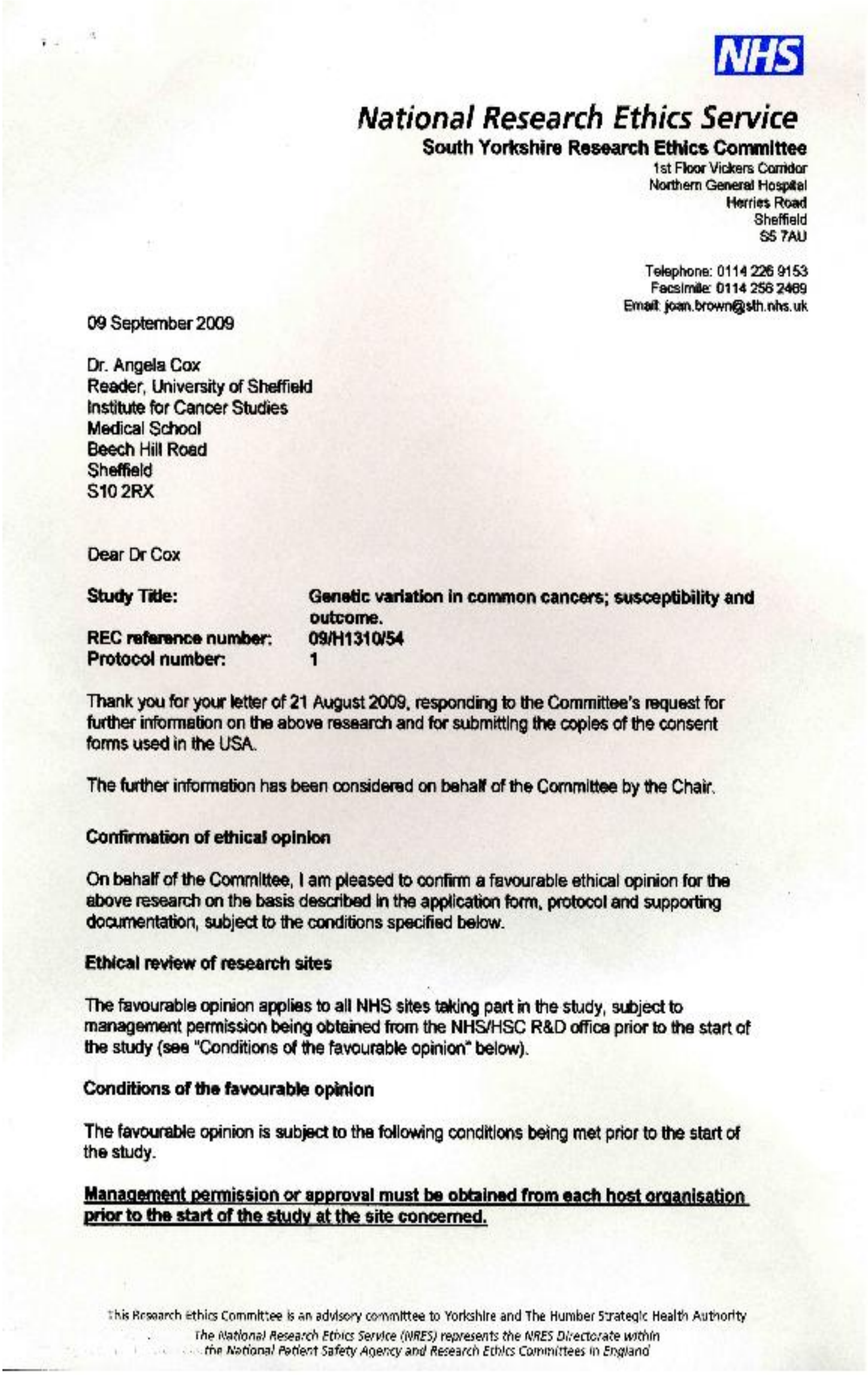
Primer	Primer Sequence 5'-3'	Length	GC%	Tm º C
*CRABP1*F	GGTGGATGTTTTGATGTAGAATTGA	25	45.6	58.5
*CRABP1*R*	**B-**ACCCAAACCTATAATACACCCTAC	24	52.4	59.8
*CRABP1*seq	ATGTTTTGATGTAGAATTGATT	22	52.4	59.8
*NEUROG1*F	AGGGAGTTTTTAGGTAGTGAAATAGAG	27	52.4	59.8
*NEUROG1*R*	**B-**ACTCCAAATACCCTCCAAATTTC	23	52.4	59.8
*NEUROG1*seq	AGGTAGTGAAATAGAGG	17	52.4	59.8

* Biotin labelled primers

Appendix 2. MS-MLPA probes- ME042-A1 Kit

Expected Length (nt)	Observed Length (nt)	Probe	Hha1 Site	Chromosomal Position
130	126.53	Reference-1		8q24
141	137.56	*IGF2*-1	+	11p15.5
148	144.28	*RUNX3*-1	+	1p36.11
160	157.41	Reference-2		10q26.13
166	163.43	*NEUROG1*-1	+	5q31.1
171	168.62	*IGF2*-2	+	11p15.5
176	174.37	*MLH1*-1	+	3p22.2
183	180.19	*CDKN2A*-1	+	9p21.3
190	186.61	Reference-3		2p22.3
195	192.25	*CDKN2A*-2	+	9p21.3
202	198.64	*NEUROG1*-2	+	5q31.1
207	204.83	*CRABP1*-1	+	15q25.1
211	209.75	*NEUROG1*-3	+	5q31.1
218	216.23	*CACNA1G*-1	+	17q21.33
226	222.54	Reference-4		6q24.3
232	231.49	CDKN2A-3	+	9p21.3
246	244.33	*CACNA1G*-2	+	17q21.33
256	253.54	*RUNX3*-2	+	1p36.11
265	263.01	*CRABP1*-2	+	15q25.1
273	271.02	*CACNA1G*-3	+	17q21.33
282	279.24	*NEUROG1*-4	+	5q31.1
292	289.93	Reference-5		19q13.13
310	307.46	*CRABP1*-3	+	15q25.1
319	316.51	*CRABP1*-4	+	15q25.1
327	323.89	Reference-6		14q22.1
335	333.95	*CDKN2A*-4	+	9p21.3
346	343.48	*RUNX3*-3	+	1p36.11
355	352.02	*MLH1*-2	+	3p22.1
364	361.49	*NEUROG1*-5	+	5q31.1
372	370.27	*RUNX3*-4	+	1p36.11
382	380	Reference-7		10q22
389	387.21	*NEUROG1*-6	+	5q31.1
399	397.54	*SOCS1*	+	16p13.13
409	407.50	*BRAF* V600E		7q34
418	416.03	*IGF2*-3	+	11p15.5
427	424.84	Reference-8		12p13.31
436	432.9	Digestion	+	17q23.2
463	461.53	*MLH1*-3	+	3p22.1
472	471.41	Reference-9		22q12.3
481	479.71	Reference-10		3p25.3

Appendix 3. Ethics approval



NHS

National Research Ethics Service

South Yorkshire Research Ethics Committee

1st Floor Vickers Corridor
Northern General Hospital
Herries Road
Sheffield
S5 7AU

Telephone: 0114 226 9153
Facsimile: 0114 256 2469
Email: joan.brown@sth.nhs.uk

09 September 2009

Dr. Angela Cox
Reader, University of Sheffield
Institute for Cancer Studies
Medical School
Beech Hill Road
Sheffield
S10 2RX

Dear Dr Cox

Study Title:	**Genetic variation in common cancers; susceptibility and outcome.**
REC reference number:	**09/H1310/54**
Protocol number:	**1**

Thank you for your letter of 21 August 2009, responding to the Committee's request for further information on the above research and for submitting the copies of the consent forms used in the USA.

The further information has been considered on behalf of the Committee by the Chair.

Confirmation of ethical opinion

On behalf of the Committee, I am pleased to confirm a favourable ethical opinion for the above research on the basis described in the application form, protocol and supporting documentation, subject to the conditions specified below.

Ethical review of research sites

The favourable opinion applies to all NHS sites taking part in the study, subject to management permission being obtained from the NHS/HSC R&D office prior to the start of the study (see "Conditions of the favourable opinion" below).

Conditions of the favourable opinion

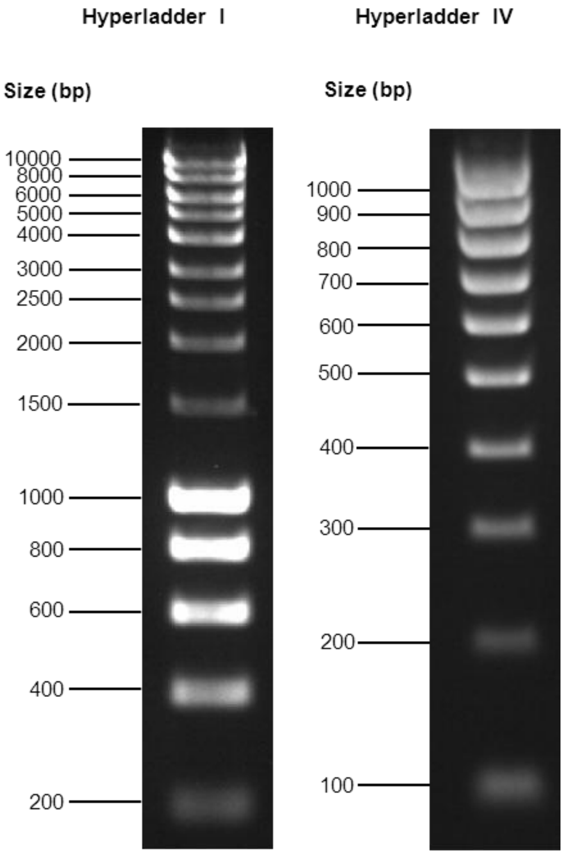
The favourable opinion is subject to the following conditions being met prior to the start of the study.

Management permission or approval must be obtained from each host organisation prior to the start of the study at the site concerned.

This Research Ethics Committee is an advisory committee to Yorkshire and The Humber Strategic Health Authority

The National Research Ethics Service (NRES) represents the NRES Directorate within
the National Patient Safety Agency and Research Ethics Committees in England

Appendix 4. DNA Size Standards



1.5% Agarose gel, 120V, 45min

Appendix 5. MS-MLPA QC metrics

Table 1 Normal DNA QC metrics results

Sample	DNA Conc./Ligation	Denaturation	Hybridisation	DNA Quant./Lig./Denat.	Sample	DNA Conc./Ligation	Denaturation	Hybridisation	DNA Quant./Lig./Denat.
N016	9.52	140.06	71.40	42.35	N016D	4.80	145.17	68.88	40.54
N023	9.96	131.67	75.95	41.08	N023D	6.60	114.78	87.12	39.31
N037	10.08	139.13	71.87	77.18	N037D	7.89	182.57	54.77	60.14
N045	7.35	142.18	70.34	108.03	N045D	7.70	194.89	51.31	99.66
N046	5.97	115.54	86.55	118.86	N046D	12.34	199.51	50.12	99.60
N053	6.95	130.13	76.84	61.62	N053D	4.24	126.96	78.77	66.41
N079	4.67	100.63	99.37	85.17	N079D	7.31	147.68	67.72	75.21
N080-1	12.97	151.53	66.00	111.94	N080D-1	13.53	184.42	54.22	82.48
N080-2	17.91	182.17	54.89	91.58	N080D-2	13.53	184.42	54.22	82.48
N083	6.04	126.26	79.20	81.32	N083D	6.80	136.81	73.09	76.26
N085	7.27	157.17	63.63	40.74	N085D	5.47	139.20	71.84	42.17
N086	8.67	85.28	117.26	106.67	N086D	9.82	106.98	93.48	84.75
N088	11.35	164.84	60.67	73.52	N088D	4.74	155.36	64.37	68.38
N090	12.74	166.19	60.17	74.42	N090D	10.74	138.10	72.41	81.37
N097	16.92	54.12	184.79	121.63	N097D	7.47	101.12	98.90	101.06
N098	11.25	85.00	117.65	104.81	N098D	13.24	157.38	63.54	85.54
N104	6.57	77.02	129.83	97.35	N104D	7.71	83.36	119.96	107.90
N107	5.53	105.52	94.77	102.08	N107D	5.35	132.10	75.70	100.75
N109	3.96	161.16	62.05	211.57	N109D	2.64	228.37	43.79	157.55
N110	5.73	125.38	79.76	144.09	N110D	10.55	103.46	96.66	83.95
N112	16.04	166.63	60.01	123.25	N112D	13.53	184.42	54.22	82.48
N114-1	5.80	108.74	91.97	139.06	N114D-1	12.51	190.10	52.60	117.18
N114-2	5.65	89.53	111.69	354.06	N114D-2	6.98	141.24	70.80	113.93
N122	7.15	144.96	68.98	42.77	N122D	3.92	144.38	69.26	46.37
N135	9.75	104.33	95.85	142.43	N135D	10.36	146.24	68.38	94.13
N138	3.35	89.30	111.99	112.22	N138D	3.67	131.65	75.96	94.68
N142	6.73	93.23	107.26	84.29	N142D	8.60	136.65	73.18	73.23
N150	10.83	161.50	61.92	48.57	N150D	4.29	158.80	62.97	59.12
N153	7.56	108.93	91.80	55.86	N153D	4.96	153.25	65.25	58.55
N158	7.64	158.61	63.05	51.19	N158D	6.43	123.97	80.67	49.80
N167	6.84	81.16	123.22	131.58	N167D	5.22	128.31	77.93	91.25
N184-1	8.01	160.69	62.23	71.75	N184D-1	7.05	171.02	58.47	68.83
N184-2	9.95	154.78	64.61	83.96	N184D-2	8.55	117.68	84.98	113.56
N201	4.68	132.05	75.73	111.96	N201D	5.84	181.32	55.15	102.37
N202	7.00	121.34	82.41	151.41	N202D	9.51	183.30	54.56	114.36
N203	6.46	111.54	89.66	115.66	N203D	7.49	98.86	101.15	99.35
N206	3.65	101.87	98.16	118.28	N206D	5.26	131.24	76.20	101.33

N208	5.40	140.36	71.25	89.42	N208D	6.51	158.17	63.22	84.76
N212	6.82	180.46	55.41	80.75	N212D	5.95	176.08	56.79	76.59
N213	12.01	136.65	73.18	87.88	N213D	5.26	136.73	73.14	83.24
N214	8.33	160.39	62.35	174.09	N214D	13.71	158.25	63.19	124.56
N218	8.74	136.08	73.49	101.40	N218D	9.33	160.34	62.37	91.77
N221-1	6.70	104.97	95.27	84.39	N221D-1	11.74	124.92	80.05	62.65
N221-2	6.70	104.97	95.27	84.39	N221D-2	11.74	124.92	80.05	62.65
N223	4.17	97.04	103.06	125.69	N223D	10.04	158.93	62.92	91.90
N244	5.40	126.45	79.08	68.59	N244D	1.43	154.85	64.58	68.74
N248	10.35	116.44	85.88	70.72	N248D	3.56	148.44	67.37	64.81
N249	6.55	138.33	72.29	63.38	N249D	5.80	149.24	67.01	62.27
N271-1	7.37	118.69	84.25	76.08	N271D-1	4.03	143.64	69.62	57.95
N271-2	7.56	188.23	53.13	93.03	N271D-2	5.75	108.67	92.02	97.88
N632	7.38	139.80	71.53	46.01	N632D	7.26	131.21	76.21	45.85
N741	5.05	124.06	80.61	53.83	N741D	8.53	105.44	94.84	74.49
N795	6.39	90.31	110.74	52.38	N795D	8.42	132.44	75.51	47.09
N824	9.17	152.41	65.61	39.09	N824D	7.86	141.96	70.44	37.32
N828	8.11	73.92	135.29	56.41	N828D	6.95	130.08	76.87	43.55
N863	3.10	85.32	117.20	64.40	N863D	2.72	133.94	74.66	63.53
N1120	32.07	73.89	135.34	74.81	N1120D	4.10	137.57	72.69	54.59
N1350	10.59	150.13	66.61	63.31	N1350D	6.41	127.17	78.63	55.50

Table 2 Tumour DNA QC metrics results

Sample	DNA Conc./Ligation	Denaturation	Hybridisation	DNA Quant./Lig./Denat.	Sample	DNA Conc./Ligation	Denaturation	Hybridisation	DNA Quant./Lig./Denat.
T016	29.31	164.88	60.65	43.43	T016D	29.31	164.88	60.65	43.43
T023	6.97	227.34	43.99	60.83	T023D	6.97	227.34	43.99	60.83
T037	27.51	173.41	57.67	60.51	T037D	14.58	174.70	57.24	53.00
T045	18.11	125.55	79.65	91.61	T045D	22.62	178.67	55.97	84.46
T046	9.39	107.76	92.80	94.20	T046D	8.18	164.90	60.64	79.22
T053	8.17	125.71	79.55	73.51	T053D	7.99	131.81	75.87	67.90
T079	7.68	102.69	97.38	87.89	T079D	8.81	135.05	74.05	66.24
T080-1	8.76	86.86	115.12	130.53	T080D-1	7.23	119.79	83.48	69.04
T080-2	8.91	163.32	61.23	92.06	T080D-2	7.23	119.79	83.48	69.04
T083	12.96	151.92	65.83	78.44	T083D	8.06	131.37	76.12	72.12
T085	7.76	190.48	52.50	48.62	T085D	6.03	171.63	58.27	49.51
T086	8.96	104.62	95.59	98.70	T086D	13.04	88.21	113.36	78.13
T088	11.28	142.66	70.10	68.40	T088D	11.55	136.03	73.52	63.68
T090	10.05	212.56	47.04	111.62	T090D	10.97	184.65	54.16	92.28
T097	12.30	79.04	126.52	91.44	T097D	9.57	94.63	105.68	80.01
T098	13.66	78.86	126.80	97.36	T098D	12.17	146.19	68.41	98.67
T104	8.18	104.05	96.11	145.31	T104D	5.18	123.06	81.26	154.05
T107	18.09	64.90	154.07	80.96	T107D	18.35	99.35	100.66	81.02

T109	12.63	54.74	182.69	89.34	T109D	12.15	77.69	128.72	82.06
T110	7.72	89.35	111.93	100.89	T110D	6.71	128.42	77.87	93.35
T112	24.47	174.30	57.37	109.36	T112D	5.80	111.91	89.35	69.78
T114-1	18.38	164.98	60.61	94.30	T114D-1	19.97	170.10	58.79	121.15
T114-2	12.73	134.37	74.42	72.35	T114D-2	9.94	109.26	91.52	76.93
T122	5.50	134.86	74.15	53.89	T122D	7.99	146.93	68.06	45.75
T135	24.37	36.11	276.90	197.46	T135D	14.59	138.29	72.31	96.89
T138	12.18	70.85	141.15	86.44	T138D	10.63	83.11	120.32	75.47
T142	7.31	195.00	51.28	150.67	T142D	5.37	147.09	67.98	154.77
T150	5.57	159.42	62.73	54.46	T150D	7.14	166.51	60.06	59.14
T153	6.64	121.86	82.06	60.14	T153D	3.67	114.86	87.07	55.10
T158	7.81	156.36	63.96	50.92	T158D	9.34	142.19	70.33	51.61
T167	5.22	79.79	125.33	100.53	T167D	12.50	105.15	95.10	76.89
T184-1	12.33	115.61	86.50	63.75	T184D-1	11.53	154.32	64.80	64.12
T184-2	7.32	165.99	60.24	65.60	T184D-2	4.34	139.41	71.73	81.72
T201	16.63	92.99	107.54	93.03	T201D	35.45	119.76	83.50	76.30
T202	6.60	78.09	128.06	96.88	T202D	6.29	100.18	99.82	78.99
T203	13.91	70.21	142.43	97.90	T203D	9.96	103.26	96.84	88.44
T206	19.06	97.90	102.15	90.66	T206D	16.78	108.05	92.55	70.33
T208	9.54	154.68	64.65	70.94	T208D	9.80	162.99	61.35	75.80
T212	7.98	132.81	75.29	58.14	T212D	5.62	140.48	71.19	60.97
T213	34.13	89.37	111.89	67.32	T213D	24.30	127.39	78.50	67.41
T214	4.96	92.15	108.51	101.58	T214D	6.60	129.67	77.12	94.04
T218	18.27	109.72	91.14	88.56	T218D	14.01	128.48	77.83	102.69
T221-1	14.16	107.42	93.09	60.30	T221D-1	11.99	121.46	82.33	66.60
T221-2	14.16	107.42	93.09	60.30	T221D-2	11.99	121.46	82.33	66.60
T223	14.06	127.61	78.37	124.12	T223D	13.58	141.16	70.84	104.44
T244	9.87	112.38	88.98	68.30	T244D	5.63	88.16	113.44	56.22
T248	10.51	116.52	85.82	65.44	T248D	10.52	147.38	67.85	60.91
T249	12.33	124.49	80.33	62.77	T249D	10.63	124.87	80.08	70.27
T271-1	9.08	143.73	69.58	72.71	T271D-1	8.33	226.34	44.18	76.60
T271-2	9.53	69.90	143.07	177.49	T271D-2	6.92	170.41	58.68	82.63
T632	14.02	144.95	68.99	51.06	T632D	4.54	143.21	69.83	49.30
T741	8.35	150.88	66.28	46.23	T741D	7.02	161.54	61.90	53.20
T795	5.72	144.58	69.17	51.44	T795D	3.60	119.00	84.04	54.82
T824	6.88	119.93	83.38	53.52	T824D	7.29	132.11	75.70	49.15
T828	8.05	222.22	45.00	50.61	T828D	7.79	146.47	68.28	58.43
T863	5.53	118.77	84.20	42.50	T863D	8.83	131.84	75.85	48.89
T1120	9.10	184.01	54.34	57.16	T1120D	4.00	149.01	67.11	58.85
T1350	8.09	139.67	71.60	51.01	T1350D	6.15	154.51	64.72	49.22

The QC metrics are described in section 2.3.4.2.1. A suffix of D represents digested DNA samples. Cells highlighted in red represent failed QC metrics.

Appendix 6. Molecular characteristics of the 53 CRC cases

Sample	CIMP	CIN	*APC* MCR	*TP53*	*KRAS*	*BRAF*	*PIK3CA*
CA016	CIMP-L	4	WT	WT	WT	WT	WT
CA023	CIMP-L	0	c.4328_4329insC	WT	WT	WT	c.1633G>A
CA037	CIMP-L	0	c.3916G>T	WT	c.35G>T	WT	WT
CA045	CIMP-L	260	c.3933_3934insT	c.524G>A	WT	WT	WT
CA046	CIMP-L	0	WT	WT	c.35G>A	WT	c.1633G>A
CA053	CIMP-L	88	WT	c.592G>T	WT	WT	WT
CA079	CIMP-N	71	c.4054_4062dup7	c.844C>T	WT	WT	WT
CA080	CIMP-L	38	c.4271_4280del10	WT	c.35G>A	WT	WT
CA083	CIMP-L	10	WT	WT	WT	WT	WT
CA085	CIMP-L	43	c.3964G>T	c.916C>T	WT	WT	WT
CA086	CIMP-L	3	WT	c.818G>A	c.35G>A	WT	WT
CA088	CIMP-H	45	c.4582_4603del22	c.220del1	c.35G>T	WT	WT
CA090	CIMP-L	63	WT	c.559+1G>T	WT	WT	WT
CA097	CIMP-L	43	c.3982C>T	c.637C>T	c.34G>T	WT	WT
CA098	CIMP-L	93	c.4037C>G	c.524G>A	WT	WT	WT
CA104	CIMP-L	37	c.3916G>T	c.734G>A	c.35G>T	WT	WT
CA107	CIMP-L	110	WT	WT	c.35G>A	WT	c.1633G>A
CA109	CIMP-H	18	c.4350C>T	WT	c.35G>T	WT	c.3140A>G
CA110	CIMP-L	0	WT	WT	WT	WT	WT
CA112	CIMP-L	1	c.4421delC	WT	WT	WT	WT
CA114	CIMP-H	74	c.4666_4667insA	WT	c.38G>A	WT	WT
CA122	CIMP-N	18	WT	WT	WT	WT	WT
CA135	CIMP-L	14	WT	WT	c.38G>A	WT	WT
CA138	CIMP-H	24	c.4127_4128delAT	WT	c.35G>C	WT	WT
CA142	CIMP-L	219	WT	c.725G>T	WT	WT	WT
CA150	CIMP-L	5	c.4216C>T	WT	WT	WT	WT
CA153	CIMP-L	150	c.4350C>T	c.844C>G	c.34G>T	WT	c.1633G>A
CA158	CIMP-N	87	WT	c.722C>G	WT	WT	WT
CA167	CIMP-L	14	WT	WT	WT	WT	WT
CA184	CIMP-L	12	c.4303A>T	c.844C>T	WT	WT	WT
CA201	CIMP-L	23	WT	WT	c.35G>T	WT	WT
CA202	CIMP-L	33	WT	WT	WT	WT	WT
CA203	CIMP-L	67	WT	WT	WT	WT	WT
CA206	CIMP-L	50	c.3883G>T	c.455-456ins1	WT	WT	WT
CA208	CIMP-L	35	c.4666_4667insA	c.730G>A & c.844C>T	c.35G>T	A1742G	WT
CA212	CIMP-N	29	WT	WT	WT	WT	WT
CA213	CIMP-L	33	WT	c.751A>C	WT	WT	WT
CA214	CIMP-L	86	WT	WT	WT	WT	WT
CA218	CIMP-L	14	c.3927_3931del5	c.734G>A	WT	WT	WT
CA221	CIMP-N	39	c.4415delT	WT	WT	WT	WT
CA223	CIMP-H	76	WT	c.811G>A	WT	WT	WT
CA244	CIMP-L	411	c.4271delC	c.817C>T	WT	WT	WT
CA248	CIMP-N	0	WT	WT	WT	WT	WT
CA249	CIMP-L	61	c.4033G>T	c.524G>A	c.35G>T	WT	WT
CA271	CIMP-H	66	WT	c.844C>T	WT	T1796A	WT
CA632	CIMP-L	18	WT	c.733G>A	c.38G>A	WT	WT
CA741	CIMP-L	66	c.4350C>T	WT	WT	WT	WT
CA795	CIMP-L	2	c.4391_4394delAGAG	WT	WT	WT	c.1633G>A
CA824	CIMP-L	21	WT	WT	c.35G>C	WT	WT
CA828	CIMP-L	248	c.3871C>T	c.524G>A	WT	WT	WT
CA863	CIMP-L	114	c.4391_4394delAGAG	c.586C>T	WT	WT	WT
CA1120	CIMP-L	2	WT	c.847_859del13	WT	WT	WT
CA1350	CIMP-L	94	WT	WT	c.34G>T	WT	WT