Genetic Mechanisms in Colorectal Polyposis

Emma L. Short
2018

A thesis submitted in candidature for the degree of Doctor of Philosophy

Division of Cancer and Genetics
School of Medicine
Cardiff University

Declarations

This work has not been submitted in substance for any other degree or award at this or any other university or place of learning, nor is being submitted concurrently in candidature for any degree or other award.

Signed……(candidate) Date……

This thesis is being submitted in partial fulfilment of the requirements for the degree of PhD.

Signed……(candidate) Date……

This thesis is the result of my own independent work/ investigation, except where otherwise stated, and the thesis has not been edited by a third party beyond what is permitted by Cardiff University's policy on the Use of Third Party Editors by Research Degree Students. Other sources are acknowledged by explicit references. The views expressed are my own.

Signed……(candidate) Date……

I hereby give consent for my thesis, if accepted, to be available online in the University's Open Access repository and for inter-library loan, and for the title and summary to be made available to outside organisations.

Signed……(candidate) Date……

Acknowledgements

First, I want to thank the patients, relatives and healthy volunteers who have taken part in this study. Without you, this work would not have been possible. I really appreciate the time you have taken to talk to me and the blood/ tissue samples you have donated. I sincerely hope that the results from this study will have a positive impact on you, your families and future generations.

I'd like to thank my supervisors Professor Julian Sampson, Dr. Karen Reed and Dr. Laura Thomas for giving me the opportunity to do this project. Thank you for your support and for teaching me the skills and techniques that have allowed me to become a scientist as well as a clinician.

Thank you to The Pathological Society of Great Britain and Ireland for giving me the grants that have allowed me to purchase necessary equipment and consumables.

Thank you to the Inherited Tumour Syndromes Research Group: to Iris and Hala for recruiting patients and obtaining blood samples; to Hannah for her suggestions about my thesis write-up and to Elena: you are a fantastic scientist and I really appreciate your help and advice.

Thanks go to Professor Geraint Williams – I've never come across someone so fiercely intelligent and yet also so kind, helpful and approachable. You're my inspiration for becoming an academic pathologist. Thank you for all your encouragement, not just during this project but over the 10 years that I have been in Cardiff.

Thanks also go to Drs. Daniel Dancer and Claudia Consoli for their help with qPCR; to Professor Ian Tomlinson, Dr. Claire Palles and Laura Chegwidden for making me so welcome in their laboratory group; to Dr. Ian Frayling for his advice about the mismatch repair genes; to Dr. Pierre Rizkallah for his help with crystal modelling of proteins; to the All Wales Medical Genetics Service, especially Justyna Tull for arranging DNA extractions and Nelson Cottrim for doing the somatic *APC* screening; to Dr. Andy Tee for his help with Western blotting; to Dr. Matthew Mort

for his assistance with statistics and to all the staff in the Cardiff and Vale Cellular Pathology Department for making so many slides for me.

I'd like to thank our international collaborators/ advisors Professor Eric Fearon, Professor Raquel Seruca and AmbryGen. Thank you for being so helpful, and for sharing information and resources which have allowed the care of the patient to be the number one priority.

Huge thanks go to Professor Trevor Dale, Dr. Anika Offergeld, Dr. Ken Ewan, Eider Valle-Encinas and all of the Dale laboratory group for allowing me to join them for the final phase of my project. Their help with the *AXIN2* functional studies has been invaluable, and I am extremely grateful for their time and support.

Most importantly, I'd like to give my heartfelt thanks to all my wonderful family for a lifetime of love, support, encouragement and inspiration. I wouldn't be where I am today without you, and I am truly grateful for everything. Finally, I want to say a special thank you to Steve, Bella and Bea, the most incredible husband and children anyone could wish for. Thank you for being my everything and for filling every day with happiness and joy. You mean the world to me.

Summary

Individuals with >/= 10 colorectal adenomas have traditionally been referred for genetic testing to identify *APC* and *MUTYH* mutations which cause Familial Adenomatous Polyposis (FAP) and *MUTYH*-Associated Polyposis (MAP) respectively. Mutations are found in most patients with >100 adenomas but in only a minority of those with 10-100 adenomas. The research described in this thesis focuses on polyposis patients with 'no mutation identified' (NMI).

The aim of this project was to identify novel genetic mechanisms causing polyposis in a cohort of 60 unrelated NMI patients. Genetic variants were sought outside of the open reading frames of *APC* and *MUTYH*, at a low frequency in *APC*, in other known and candidate 'polyposis genes' and in whole exomes. Novel variants were characterised genetically and functionally to provide evidence for or against their clinical significance.

Rare variants were found in the 5'UTR of *APC*, the mismatch repair (MMR) genes, *POLE*, *POLD1* and *AXIN2*. The 5'UTR *APC* variant, c.-190A>G, was associated with reduced transcript levels and segregated with FAP in a multiplex family, allowing translation to diagnostic testing. Three additional patients had reduced *APC* transcript levels, but the cause was not determined by deep sequencing of their genomic *APC* loci. The MMR gene variants were deemed unlikely to be pathogenic as associated tumours were microsatellite stable with normal MMR protein immunohistochemistry. Studies into the *pol* gene variants are ongoing. The *AXIN2* mutation, c.1642G>T, p.Glu548*, was identified in a family with polyposis and ectodermal dysplasia. Most of their adenomas appeared to lack *APC* mutations and *in vitro* studies suggested that the mutation may impair inhibition of Wnt-signalling.

Gene panel testing using next-generation sequencing technologies may improve molecular genetic diagnosis of previously NMI patients but additional characterisation of novel variants is likely to be required for clinical translation, with the ultimate aim of preventing colorectal cancer.

Table of Contents

Appendices are available in the attached CD.

List of Figures

List of Tables

Abbreviations

°C	Degrees Celsius
µ	Micro
µg	Microgram
µl	Microlitre
µm	Micrometre
8-oxodG	8-hydroxyguanine
AD	Autosomal Dominant
AFAP	Attenuated Familial Adenomatous Polyposis
AI	Allelic Imbalance
APC	Adenomatous Polyposis Coli
AR	Autosomal Recessive
AWMGS	All Wales Medical Genetics Service
BCA	Bicinchoninic Acid
BER	Base Excision Repair
BMP	Bone Morphogenetic Protein
Bp	Base Pair
BR	Broad Range
BSA	Bovine Serum Albumin
CADD	Combined Annotation Dependent Completion
cDNA	Complementary Deoxyribonucleic Acid
CHRPE	Congenital Hypertrophy of the Retinal Pigment Epithelium
CIMP	CpG Island Methylator Phenotype
CIN	Chromosomal Instability
COOT	Crystallographic Object-Oriented Toolkit
CRC	Colorectal Cancer
dAMP	Deoxyadenosine Monophosphate
dbSNP	Single Nucleotide Polymorphism Database
DDD	Digital Differential Display
DGGE	Denaturing Gradient Gel Electrophoresis
DHPLC	Denaturing High Performance Liquid Chromatography
DNA	Deoxyribonucleic Acid
dNTP	Deoxyribose Nucleoside Triphosphate
ds	Double Stranded

Dsh	Dishevelled
EB1	End Binding Protein 1
EDTA	Ethylenediaminetetraacetic Acid
FAP	Familial Adenomatous Polyposis
FBS	Fetal Bovine Serum
FFPE	Formalin Fixed Paraffin Embedded
gDNA	Genomic DNA
GDP	Guanosine Diphosphate
GSK3β	Glycogen Synthase Kinase 3β
GTP	Guanosine Triphosphate
HA	Heteroduplex Analysis
HF	High Fidelity
HGD	High Grade Dysplasia
HGMD	Human Gene Mutation Database
HMPS	Hereditary Mixed Polyposis Syndrome
HP/ HPP	Hyperplastic Polyp
HPS	Hyperplastic Polyposis Syndrome
HS	High Sensitivity
IGV	Integrative Genomics Viewer
IMG	Institute of Medical Genetics
IVSP	In Vitro Synthesised Protein Assay
JPS	Juvenile Polyposis Syndrome
k	Kilo
Kb	Kilobase
kDa	Kilodalton
LB	Lysogeny Broth
LDS	Lithium Dodecyl Sulphate
LEF	Lymphoid Enhancer Factor
LGD	Low Grade Dysplasia
LOH	Loss of Heterozygosity
LS	Lynch Syndrome
M	Milli
MAP	MUTYH-Associated Polyposis
MCR	Mutation Cluster Region
MES	2-(N-morpholino) ethanesulphonic acid
miRNA	Micro Ribonucleic Acid

ml	Millilitre
MLPA	Multiplex Ligation-Dependent Probe Analysis
MMG	Megamix Gold
MMR	Mismatch Repair
M-PVA	Polyvinyl Alcohol Particle Magnetic Beads
mRNA	Messenger Ribonucleic Acid
MSI	Microsatellite Instability
MSP	Methylation Specific PCR
n	Nano
NER	Nucleotide Excision Repair
NFQ	Non-Fluorescent Quencher
Ng	Nanogram
NGS	Next Generation Sequencing
NHS	National Health Service
NMD	Nonsense Mediated Decay
NMI	No Mutation Identified
NTC	No Template Control
ORF	Open Reading Frame
PBS	Phosphate Buffered Saline
PCR	Polymerase Chain reaction
PDB	Protein Databank
PI3K	Phosphatidylinositol-4,5-bisphosphate 3-kinase
PJS	Peutz-Jegher's Syndrome
Poly	Polymorphism
PPAP	Polymerase Proofreading Associated Polyposis
PTT	Protein Truncation Test
Px	Patient
QC	Quality Control
qPCR	Quantitative Polymerase Chain Reaction
Rcf	Relative Centrifugal Force
REC	Research Ethics Committee
RNA	Ribonucleic Acid
RNase	Ribonuclease
RNAsin	Ribonuclease Inhibitor
ROS	Reactive Oxygen Species
Rpm	Revolutions Per Minute

RT	Reverse Transcription
S	Seconds
SDM	Site Directed Mutagenesis
SDS	Sodium Dodecyl Sulphate
smMIPs	Single Molecule Molecular Inversion Probes
SNP	Single Nucleotide Polymorphism
SOC	Super Optimal Broth with Catabolite Repression
SOP	Standard Operating Procedure
SPS	Serrated Polyposis Syndrome
ss	Single Stranded
SSA	Sessile Serrated Adenoma
SSCP	Single Strand Conformation Polymorphism
SSP	Sessile Serrated Polyp
TA	Tubular Adenoma
Ta	Annealing Temperature
TAE	Tris Base, Acetic Acid and Ethylenediaminetetraacetic Acid
TBS	Tris-Buffered Saline
TC	Tissue Culture
TCF	T Cell Factor
TE Buffer	Tris EDTA Buffer
TGFβ	Transforming Growth Factor β
Tm	Melting Temperature
TSA	Traditional Serrated Adenoma
TVA	Tubulovillous Adenoma
U	Units
UCSC	University of California Santa Cruz
UDS	Ultra-Deep Sequencing
UNG	Uracil-N-Glycosylase
UTR	Untranslated Region
UV	Ultraviolet
V	Volts
VLB	Variant Likely Benign
VLP	Variant Likely Pathogenic
VUS	Variant of Unknown Significance
VA	Villous Adenoma
WGP	Wales Gene Park

Chapter 1 General Introduction

1.1 Background

In 2016, cancer was the leading cause of mortality in England and Wales, accounting for 28.5% of all deaths (Office for National Statistics, 2017). It is estimated that the lifetime risk of developing any malignant neoplasm, excluding non-melanoma skin cancer, is 40% for males and 37% for females (Sasieni *et al* 2011). Cancer refers to the process in which cells grow uncontrollably. The transformation of a normal cell into a malignant one is a multistep process, in which several genetic alterations occur, allowing the cell to escape from normal control mechanisms (Hanahan and Weinberg, 2000). The genetic changes involve the activation of oncogenes and the silencing of tumour suppressors (Hanahan and Weinberg, 2011). These changes allow the cell to attain a growth advantage, for example through self-sufficiency in growth signals, insensitivity to growth inhibitory signals, evasion of apoptosis, limitless replicative potential, sustained angiogenesis, tissue invasion/ metastasis, reprogramming of cellular energy metabolism and immune evasion (Hanahan and Weinberg, 2011; Hanahan and Weinberg, 2000).

In order to acquire these advantageous traits, there is a succession of alterations in the genomes of neoplastic cells (Hanahan and Weinberg, 2011). Mutations may occur spontaneously or may be triggered by external factors such as viruses, radiation and chemicals. In addition to mutation events, gene expression can be altered through epigenetic mechanisms such as DNA methylation, histone modification and micro-RNA (miRNA) expression (Berdasco and Esteller 2010; reviewed in Jones and Baylin 2007). Ultimately, most cancers develop as a result of a complex interaction between an individual's genetic make-up and their environment.

1.2 Colorectal Carcinoma

Colorectal carcinoma (CRC) is the fourth most common cancer in the United Kingdom (UK) and the second most common cause of cancer death (Cancer Research UK 2015). Between 2013- 2015, approximately 41 700 people were diagnosed with CRC, which is more than 110 people every day (Cancer Research UK 2015). An individual's lifetime risk of developing CRC is 5%, but this figure

increases dramatically with age (reviewed in Fodde 2002). The incidence of CRC is generally high in developed countries, with a 20-fold difference in incidence rates between high- and low-risk geographical areas (reviewed in Fodde 2002). The difference is thought to largely result from environmental factors, in particular differences in diet (Fodde 2002).

The majority of CRCs occur sporadically, but in 15-35% of patients, hereditary factors are important (reviewed in Mishra and Hall 2012; Burt 2007). In approximately 5% of cases, the disease is caused by a highly penetrant dominantly inherited syndrome (reviewed in Mishra and Hall 2012). The most common is Lynch Syndrome, due to inherited defects in the mismatch repair (MMR) system. This accounts for 2-5% of cases of CRC (reviewed in Mishra and Hall 2012). Familial Adenomatous Polyposis (FAP), due to germline mutations in the *APC* gene, is responsible for <1% of the disease burden and non-syndromic familial presentations comprise 20% of cases (reviewed in Mishra and Hall 2012).

CRCs result from the progressive accumulation of genetic and epigenetic alterations which cause normal colonic epithelium to transform into adenocarcinoma (Grady and Carethers 2008). Recent work based on gene expression profiling (Guinney *et al* 2015) suggests that CRC can be classified into 4 molecular subtypes: MSI Immune, Canonical, Metabolic and Mesenchymal. However, the traditional approach has been to categorise tumours into 3 groups: those with chromosomal instability (CIN), those with microsatellite instability (MSI) and those with a hypermethylated phenotype (CpG Island Methylator Phenotype or 'CIMP'). There is considerable overlap between the latter two groups.

1.2.1 Chromosomal Instability (CIN), *APC* and the Wnt-Pathway

The vast majority of CRCs develop from pre-existing adenomas. Such tumours are characterised by chromosomal instability, which is seen in 80-85% of colorectal tumours (reviewed in Grady and Carethers 2008). There are certain key genetic mutational events which occur, allowing the progression from normal epithelium, to dysplasia, and finally invasive malignancy. The loss of adenomatous polyposis coli (*APC*) gene function seems to be the initiating event, followed by mutations in *KRAS, SMAD4* and *p53* (Fodde 2002). In keeping with Knudson's two hit hypothesis of tumourigenesis, two mutational events are required to knock out the

functioning of a tumour suppressor gene (Knudson 1971), whilst activation of an oncogene requires only one mutation.

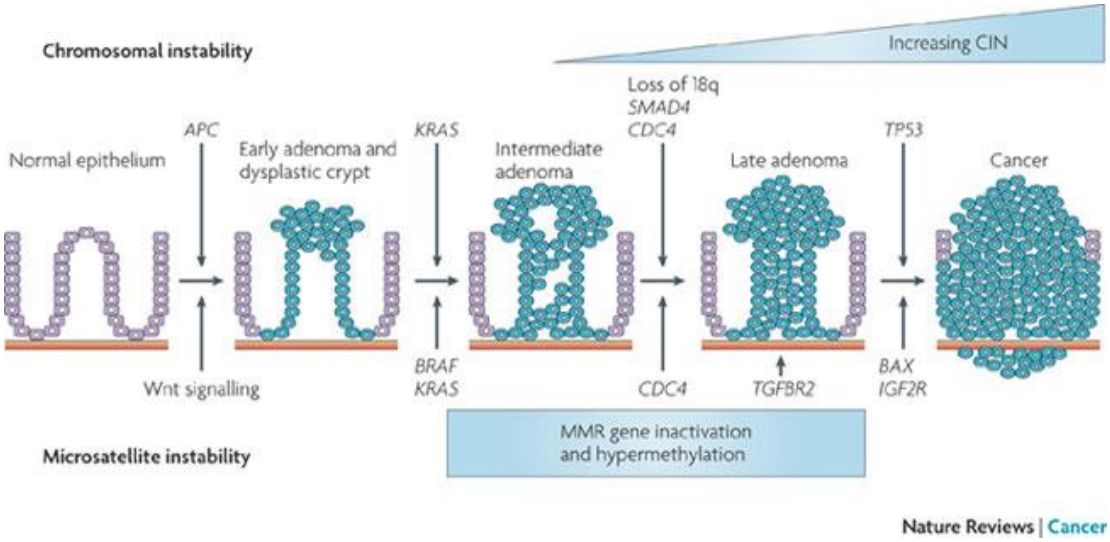


Figure 1.1: Image from Walther et al (2009) illustrating the progressive accumulation of genetic changes involved in the malignant progression of colorectal tumours

1.2.1.1 Adenomatous Polyposis Coli (*APC*)

The *APC* gene is found on chromosome 5 (5q22.2). It consists of 8535 coding base pairs, encoding a 2843 amino acid multidomain protein. Exon 15 is responsible for more than 75% of the coding sequence of the gene and is the most common site for germline and somatic mutations (reviewed in Fearnhead *et al* 2002; reviewed in Fearnhead *et al* 2001). 95% of CRC-associated *APC* mutations are nonsense or frameshift mutations, creating a truncated protein with abnormal function (Bodmer 1999).

The APC protein is a 312kDa tumour suppressor, which is involved in many cellular processes including intercellular adhesion, signal transduction, proliferation, apoptosis and migration. One of its major roles is in regulating cytoplasmic levels of β-catenin, thus negatively regulating Wnt signalling (Mishra and Hall 2012; reviewed in Fearnhead *et al* 2002; Fodde 2002; Fearnhead *et al* 2001).

1.2.1.2 The Wnt pathway

The Wnt proteins are a family of signalling proteins which are involved in developmental events during embryogenesis and in adult tissue homeostasis (Logan and Nusse 2004). They have multiple effects within a cell, including

triggering cell division, cell fate specification and differentiation (Logan and Nusse 2004).

Wnt proteins bind to Frizzled/ low density lipoprotein receptor-related proteins, which are found on cell surface membranes (Logan and Nusse 2004). This transduces a signal to intracellular proteins, including Dishevelled (Dsh), Glycogen Synthase Kinase-3β (GSK-3), AXIN, APC and β-catenin (Logan and Nusse 2004). In the absence of Wnt signalling, β-catenin levels are usually low: a complex composed of GSK-3, APC and AXIN targets it for ubiquitin-mediated degradation (Logan and Nusse 2004; reviewed in Fearnhead *et al* 2002; reviewed in Fearnhead *et al* 2001). When cells receive Wnt signals, the degradation pathway is inhibited. This allows β-catenin to accumulate in the cytoplasm and nucleus. In the nucleus, it complexes with one of the T cell factor (TCF) or lymphoid enhancer factor (LEF) transcription factors, to initiate transcription of a range of genes, including c-*myc* and *cyclin D1* (Logan and Nusse 2004; reviewed in Fearnhead *et al* 2002; reviewed in Fodde 2002; reviewed in Fearnhead *et al* 2001). *Myc* and *cyclin D1* are both relevant to tumourigenesis as they have roles in proliferation, apoptosis and cell-cycle progression (reviewed in Fodde 2002).

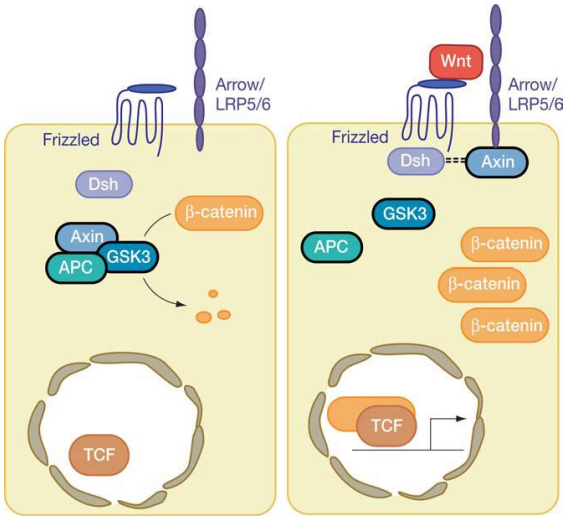


Figure 1.2: Image from Logan and Nusse (2004) illustrating the Wnt pathway in both the absence and presence of Wnt stimulating molecules

In the normal intestinal epithelium, nuclear β-catenin expression is higher in the proliferative component, and APC levels are higher in post-replicative cells (reviewed in Fodde 2002). These findings support β-catenin signalling having a role

in maintaining stem cell properties and controlling differentiation in the bowel (reviewed in Fodde 2002). As cells move along the crypt-villous axis, increasing levels of APC counteract β-catenin signalling and allow differentiation to occur (reviewed in Fodde 2002). *APC* mutations hence allow increased numbers of stem cells and reduced cellular differentiation (Fodde 2002).

In addition to their roles in the initial stages of the adenoma-carcinoma sequence, *APC* mutations remain important throughout malignant progression. Nuclear β-catenin staining strongly correlates with tumour size and dysplasia, and high levels of nuclear β-catenin have been found at the invasive fronts of adenocarcinomas (reviewed in Fodde 2002). The APC protein is also involved in chromosomal stability at mitosis: it has an EB1-binding domain in its C-terminal end, which associates with the growing ends of cytoplasmic and spindle microtubules, as well as centrosomes. *APC* mutant cells are hence characterised by chromosomal instability which is observed in the majority of CRCs (reviewed in Grady and Carethers 2008; reviewed in Fodde 2002).

APC mutations can therefore be seen to have a key role in both initiating and promoting CRC: activation of the Wnt pathway affects cell proliferation, apoptosis, and possibly differentiation of intestinal stem cells (Fodde 2002) and at later stages of carcinogenesis, CIN resulting from *APC* mutations can accelerate tumour progression (Fodde 2002).

1.2.1.3 *KRAS, SMAD4* and *P53*

As illustrated in Figure 1.1, further genes involved in colorectal tumourigenesis include *KRAS, SMAD4,* and *p53*. The importance of *ras* gene mutations in colorectal carcinogenesis was first reported in 1987 (Bos *et al* 1987; Forrester *et al* 1987). The *K-ras* oncogene has been found to be mutated in 10-15% of adenomas <1cm, and in 30-60% of adenomas >1cm and carcinomas (reviewed in Brink *et al* 2003; reviewed in Fearon and Vogelstein 1990). The gene encodes a 21kDA protein located in the inner plasma membrane, with intrinsic GTPase activity. It is involved in the transduction of mitogenic signals (reviewed in Brink *et al* 2003). It is activated by a diverse spectrum of extracellular stimuli, such as growth factors, cytokines and hormones (reviewed in Brink *et al* 2003; Shields *et al* 2000). Once activated, it stimulates a multitude of downstream signalling cascades, including the

Raf serine/ threonine kinases, phosphoinositide 3-kinases (PI3Ks) and a family of GDP-GTP exchange factors (reviewed in Shields *et al* 2000).
Mutant KRAS has impaired GTPase activity, meaning that it is constitutively active (reviewed in Brink *et al* 2003). This can cause uncontrolled cell growth and proliferation. A *KRAS* mutation in a colonic epithelial cell which already has *APC* mutations results in a clonal expansion and increased risk of progression to cancer (reviewed in Vogelstein and Kinzler 2004).

The *SMAD4* gene is on chromosome 18q. It was first identified as a tumour suppressor gene in pancreatic cancer in 1996 (Hahn *et al* 1996). *SMAD4* mediates the TGFβ signalling pathway to suppress epithelial growth (reviewed in Miyaki and Kuroki 2003). The SMAD4 protein acts as a trimer and forms complexes with additional SMAD proteins: receptor-phosphorylated SMAD2 and SMAD3 (Woodford-Richens *et al* 2001). These complexes then translocate from the cytoplasm to the nucleus and associate with DNA binding factors to facilitate the transcription of target genes, including cyclin-dependent kinase inhibitors such as p15(ink4B) and the inhibitory SMAD7 (Woodford-Richens *et al* 2001). Loss of SMAD4 function may result in the loss of transcription of genes necessary for cell-cycle control (Woodford-Richens *et al* 2001). Cells may therefore become TGF-β resistant and escape from TGF-β-mediated growth control and apoptosis (Woodford-Richens *et al* 2001). *SMAD4* is mutated in a significant proportion of colorectal tumours, with the frequency of mutational events increasing with the progression of carcinogenesis: it is mutated in 0% of adenomas, 10% of 'intramucosal carcinomas', 7% of carcinomas without metastases and 35% of carcinomas with distant metastases (reviewed in Miyaki and Kuroki 2003).

In 1988, it was reported that 73% of CRCs, 47% of 'advanced adenomas' and 11-13% of 'early stage adenomas' had a deletion of part of chromosome 17 (Vogelstein *et al* 1988). This region was subsequently shown to include the *p53* gene (17p13.1) (Baker *et al* 1989). The p53 protein is a transcription factor which has a vital role in maintaining genomic stability (reviewed in Sarasqueta *et al* 2013). Following DNA damage, p53 activation causes arrest of the cell cycle to allow DNA repair (reviewed in Sarasqueta *et al* 2013). If the damage is too extensive, p53 can drive a cell towards senescence or apoptosis (reviewed in Sarasqueta *et al* 2013; reviewed in Vogelstein and Kinzler 2004). The functional loss of p53 is a key event in the malignant progression of a colorectal adenoma to CRC (reviewed in Iacopetta 2003; Vogelstein *et al* 1988).

1.3 Microsatellite Instability (MSI)

The human mismatch repair system (MMR) involves 7 key genes: *MSH2, MSH6, MSH3, MLH1, PMS1, PMS2* and *MLH3.* Their protein products are able to recognise and repair nucleotide mismatches which have escaped the normal editing function of DNA polymerase (reviewed in Grady and Carethers 2008). If such mismatches are not repaired, nucleotide transitions or transversions result, allowing potentially oncogenic mutations to occur more frequently, leading to a 'hypermutable phenotype' (reviewed in Grady and Carethers 2008).

Lynch Syndrome (LS) is an autosomal dominant (AD) disease, which accounts for approximately 5% of cases of CRC (reviewed in Mishra and Hall 2012). Patients also have an inherited predisposition to a range of other malignancies, in particular endometrial carcinoma (Lynch *et al* 2015). LS occurs due to inherited mutations of *MSH2, MSH6, MLH1* or *PMS2* (reviewed in Grady and Carethers 2008). In 15-20% of sporadic colon cancers, inactivation of the mismatch repair (MMR) system occurs, either though methylation of *MLH1* or point mutations in *MLH1*/ *MSH2*/ other members of the MMR family (reviewed in Grady and Carethers 2008). This leads to microsatellite instability (MSI).

It is thought that certain key tumour suppressor genes drive the pathogenesis of MSI tumours, and these are different to those which are mutated in CIN tumours (reviewed in Grady and Carethers 2008). In around 85% of colorectal tumours with MSI, a repeat of 10 adenines undergoes a frameshift mutation in the *TGFBR2* gene. This allows tumour cells to escape the growth suppressing effects of TGF-β1 (reviewed in Grady and Carethers 2008). Another gene commonly mutated in MSI CRC is *BAX*, which plays a role in apoptosis. It is mutated in 50% of MSI CRC (reviewed in Grady and Carethers 2008). Interestingly, frameshift mutations within coding mononucleotide repeats are also seen in *APC* (reviewed in Lynch *et al* 2015).

MSI CRCs tend to have a certain clinical and pathological phenotype. They generally occur in the right side of the colon, and microscopically they have a mucinous appearance with large numbers of tumour-infiltrating lymphocytes (reviewed in Grady and Carethers 2008).

1.4 CpG Island Methylator Phenotype (CIMP)

Epigenetic regulation of gene expression can be achieved through methylation of CpG islands found in gene promoters, causing silencing of the downstream gene. Such silencing of tumour suppressors and/ or DNA repair genes is a common feature of human neoplasia (reviewed in Hughes *et al* 2012). Widespread CpG island promoter methylation is referred to as the CpG island methylator phenotype (CIMP) and was first described in 1999 (Toyota *et al* 1999). The cause of CIMP remains to be elucidated but may potentially result from aberrant *de novo* methylation or through the loss of protection against *de novo* methylation (reviewed in Toyota *et al* 1999). Environmental factors, such as anthropometry and physical activity, smoking and alcohol may also play a role (reviewed in Hughes *et al* 2012).

CRC which exhibit CIMP are thought to arise via the 'serrated pathway of neoplasia'. The precursor lesions are hyperplastic polyps, rather than adenomas, and an early event is a mutation of the *BRAF* oncogene (reviewed in Guarinos *et al* 2012). BRAF is a component of the MAPK signalling pathway. The pathway involves activation of cell membrane signalling molecules with subsequent stimulation of cytoplasmic protein kinases (Seger and Krebs 1995). The transmitted signals eventually activate cellular processes such as proliferation, differentiation and development (Seger and Krebs 1995). Activating mutations of BRAF increase its kinase activity, which drives the proliferation of malignant cells (reviewed in Bollag *et al* 2012).

Most CIMP CRCs have epigenetic silencing of *MLH1*, leading to microsatellite instability (reviewed in Hughes *et al* 2012), and may have silencing of tumour suppressor genes such as *p16* (Toyota *et al* 1999).

Typically, CIMP tumours are associated with older age, female sex and occurrence in the right side of the bowel (reviewed in Hughes *et al* 2012), as is seen with MSI CRC.

1.5 The Colorectal Polyposis Syndromes

Colorectal polyps are masses of tissue which are found projecting from the mucosa of the large bowel. They are classified according to their microscopic appearance, and include adenomas, hyperplastic polyps and hamartomatous polyps. Most polyps occur sporadically, but some are seen as part of a genetic 'polyposis

syndrome' (Figure 1.3). Colorectal polyps are benign but are clinically significant as they may confer a risk of malignancy.

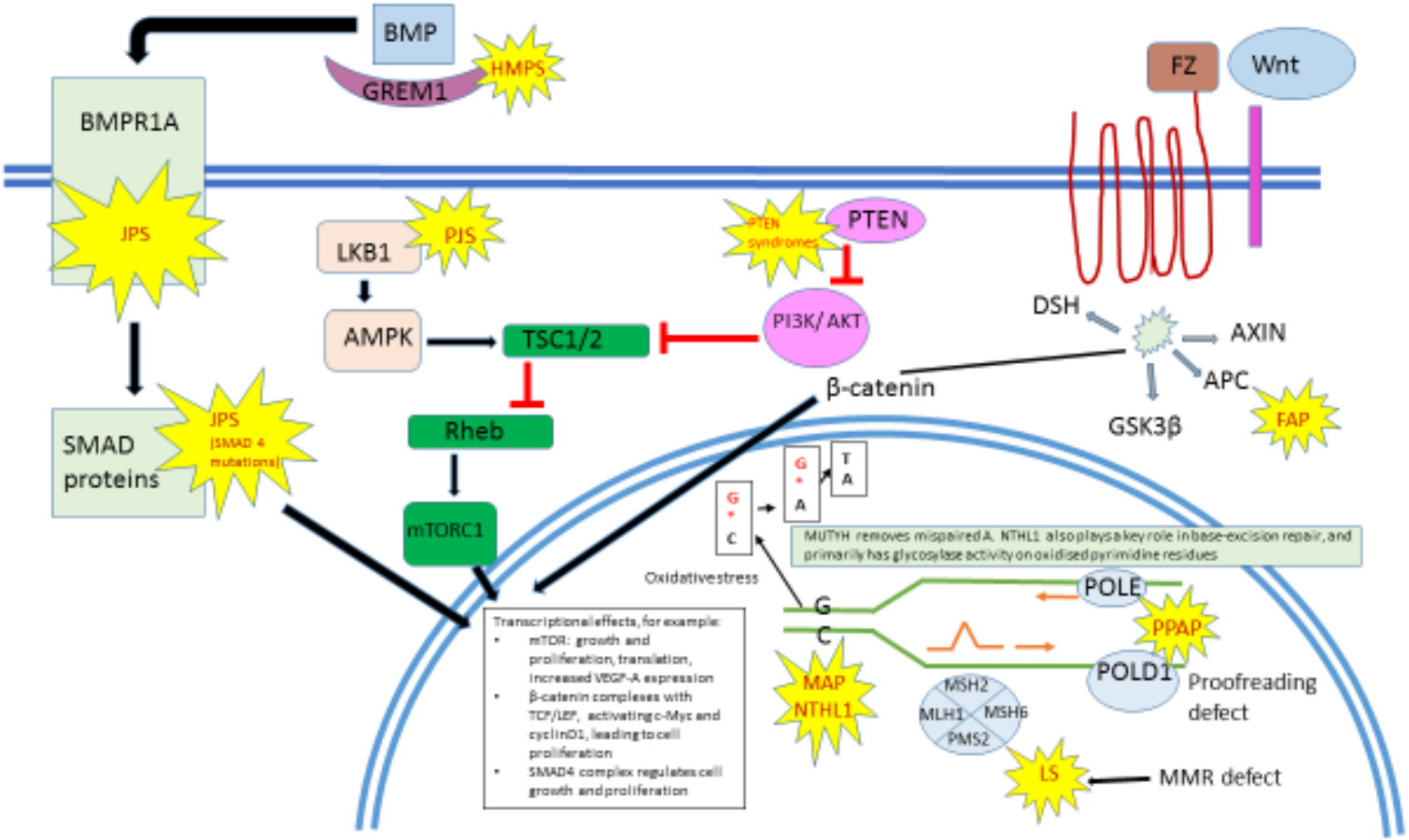


Figure 1.3: Diagram illustrating relationships between genes and signalling pathways involved in inherited colorectal polyposis syndromes. FAP, Familial Adenomatous Polyposis; HMPS, Hereditary Mixed Polyposis Syndrome; JPS, Juvenile Polyposis Syndrome; LS, Lynch Syndrome; MAP, MUTYH-Associated Polyposis; NTHL1, NTHL1-Associated Polyposis; PJS, Peutz-Jegher's Syndrome; PPAP, Polymerase Proofreading-Associated Polyposis. Image modified from Short et al (2015)

1.5.1 Colorectal Adenomas

Colorectal adenomas are common. They are found in between 1.72% and 63% of autopsies (Pendergrass *et al* 2008; Paspatis *et al* 2001; Correa *et al* 1977; Arminski and McLean 1964; Chapman 1963). In asymptomatic patients undergoing colonoscopies, adenoma prevalence is between 6.3% and 41% (Chung *et al* 2010; Rundle *et al* 2008; Lin *et al* 2006; Strul *et al* 2006; Soon *et al* 2005; reviewed in Giacosa 2004; Yamaji *et al* 2004; DiSario *et al* 1991). They occur more frequently in males than in females, and their prevalence increases with age (Chung *et al* 2010; Pendergrass *et al* 2008; Lin *et al* 2006; Strul *et al* 2006; Soon *et al* 2005; Yamaji *et al* 2004; Paspatis *et al* 2001; DiSario *et al* 1991; Correa *et al* 1977; Chapman 1963).

The significance of colorectal adenomas is that they are pre-malignant lesions. The majority of colorectal CRCs are thought to develop from pre-existing adenomas. The probability that an adenoma will become malignant depends upon its size, morphology and degree of dysplasia. Large villous lesions harbouring high grade dysplasia confer the highest risk (Terry *et al* 2002; O'Brien *et al* 1990; Shinya and Wolff 1979; Muto *et al* 1975).

Most colorectal adenomas occur sporadically. However, there are syndromes of colorectal polyposis, in which patients develop multiple polyps as a result of an underlying genetic mutation. These include Familial Adenomatous Polyposis (FAP), *MUTYH*-Associated Polyposis (MAP), Polymerase Proofreading-Associated Polyposis (PPAP), *NTHL1*-Associated Polyposis/ CRC and *MSH3*-Associated Polyposis/ CRC.

1.5.1.1 Familial Adenomatous Polyposis (FAP)

FAP is a dominantly inherited Mendelian trait, in which patients develop hundreds to thousands of colorectal adenomas during adolescence or the third decade of life (reviewed in Fearnhead *et al* 2002; Bodmer 1999). All such patients will invariably develop CRC if they are left untreated (Bodmer 1999; reviewed in Fearnhead *et al* 2002).

The first case of histologically verified adenomatous polyposis was published in 1881 by Sklifasowski (reviewed in Bülow *et al* 2006). The following year Harrison-

Cripps described ‘disseminated polypus of the rectum’ in two teenage siblings, both of whom had 20-30 colorectal polyps (reviewed in Bülow *et al* 2006; Harrison-Cripps 1882). In the late 1800s there were numerous case reports describing patients with multiple colorectal adenomas, and an association with colorectal malignancy was noted (reviewed in Bülow *et al* 2006). In 1925, Lockhart-Mummery stated that the ‘condition of multiple adenomata was invariably antecedent to carcinoma’ and that ‘the condition of multiple adenomata is often hereditary in a marked degree’ (Lockhart-Mummery 1925).

It is now known that FAP, and an attenuated form of the disease, AFAP, are due to germline or somatic mosaic mutations in *APC*. Over 1500 different mutations in *APC* have been identified to date (Kadiyska *et al* 2013). The majority of mutations are inherited. It used to be thought that approximately one quarter of cases occurred *de novo* (reviewed in Fearnhead *et al* 2001), but this is an overestimate, as this figure included apparent *de novo* patients who actually had *MUTYH*-Associated Polyposis (MAP). A third of all germline mutations occur at codons 1061 and 1309, with the remainder spread relatively uniformly between codons 200 and 1600 (reviewed in Fearnhead *et al* 2001). The nature of the germline mutation determines the nature of the second somatic hit to *APC* (reviewed in Fearnhead *et al* 2002; reviewed in Fearnhead *et al* 2001). Germline mutations occurring between codons 1194 and 1392 tend to be followed by allelic loss of *APC* as a second hit (loss of heterozygosity, LOH), whereas germline mutations lying outside of this region tend to be associated with truncating mutations in the mutation cluster region (MCR) between codons 1286 and 1513 (reviewed in Fearnhead *et al* 2002). The reason for this may be related to the resultant level or functional characteristics of APC protein produced: it is proposed that to allow efficient tumourigenesis, the function of APC must be impaired sufficiently to allow a certain level of nuclear β-catenin accumulation, but that β-catenin levels must not be too great, or this can result in apoptosis (Albuquerque *et al* 2002).

The incidence of FAP is approximately 1 per 8000, and it accounts for around 0.5% - 1% of CRC (Mishra and Hall 2012; reviewed in Fearnhead *et al* 2002; reviewed in Bodmer 1999). In addition to colorectal adenomas, FAP patients may develop extra-intestinal manifestations of their disease, for example congenital hypertrophy of the retinal pigment epithelium (CHRPE), duodenal and peri-ampullary tumours, desmoid tumours, papillary carcinoma of the thyroid, medulloblastoma,

hepatoblastoma, osteomas and epidermoid cysts (Mishra and Hall 2012; reviewed in Fearnhead *et al* 2002).

1.5.1.2 *MUTYH*-Associated Polyposis (MAP)

Prior to 2002, inherited defects of base excision had not been associated with any human genetic disorder (Al-Tassan *et al* 2002). That year, mutations in the *MUTYH* gene were shown to cause an inherited predisposition to colorectal tumours (Al-Tassan *et al* 2002).

The *MUTYH* gene is located on the short arm of chromosome 1 (1p32.1-p34.3) (Poulsen and Bisgaard 2008). It consists of 16 exons and encodes a protein of 535 amino acids, the MUTYH glycosylase (Poulsen and Bisgaard 2008). MUTYH glycosylase is part of the base excision repair (BER) system. It is involved in repairing DNA mismatches occurring as a result of oxidative DNA damage (Mazzei *et al* 2013). Each human cell metabolises approximately 10^{12} molecules of oxygen per day (reviewed in Nohmi *et al* 2005). About 1% of oxygen metabolism results in the production of reactive oxygen species (ROS) which include superoxide, hydrogen peroxide, hydroxyl radicals and singlet oxygen (reviewed in Nohmi *et al* 2005). ROS can damage DNA, producing 8-hydroxyguanine (8-oxodG). This frequently pairs with dAMP. Under normal circumstances, this mispairing would be repaired by MUTYH, to create C: 8-oxodG base pairs. Another enzyme, OGG1 will then remove the 8-oxodG. Hence the combined effects of MUTYH and OGG1 will prevent GC > TA transversions (Mazzei *et al* 2013).

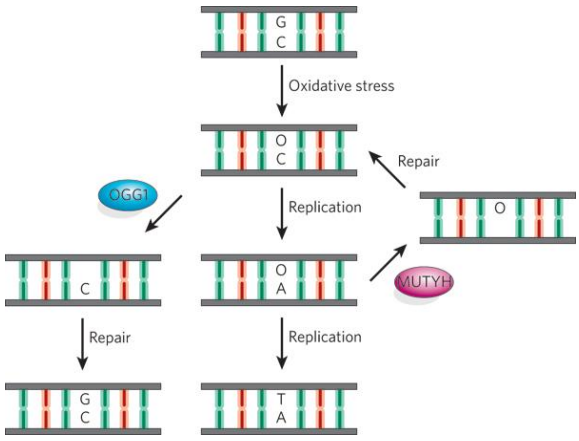


Figure 1.4: Image from David et al (2007) illustrating the actions of MUTYH *in the repair of oxidative DNA damage*

Patients with biallelic germline *MUTYH* mutations are predisposed to mutations in genes including *APC* and *KRAS.* The clinical manifestation of this is MAP. MAP is an autosomal recessive (AR) disease, in which patients develop multiple colorectal adenomas. The mean age of diagnosis is 45-50 years, and patients typically have between 10 and 100 polyps (reviewed in Mazzei *et al* 2013; reviewed in Poulsen and Bisgarrd 2008; Croitoru *et al* 2007; Nielsen *et al* 2007; Gismondi *et al* 2004; Wang *et al* 2004; Sampson *et al* 2003; Sieber *et al* 2003). Some patients do not develop polyps but present with cancer (Farrington *et al* 2005; Wang *et al* 2004; Enholm *et al* 2003). Although the majority of polyps are adenomas, hyperplastic polyps (HPPs) and sessile serrated adenomas (SSAs) are also seen (reviewed in Mazzei *et al* 2013; Boparai *et al* 2008; Lipton *et al* 2003a).

Patients with MAP have an increased risk of developing CRC. Malignancy has been reported with varying frequencies: one paper reported a prevalence of 19.47% at 50 years, and 42.89% at 60 years (Lubbe *et al* 2009); another found that 48% of patients with MAP developed CRC with a mean age of diagnosis of 49.7 years (Sampson *et al* 2003). It has been suggested that biallelic inactivation of *MUTYH* imparts an overall 93-fold excess risk and that all homozygotes/ compound heterozygotes will develop CRC by age 60 (Farrington *et al* 2005). Win *et al* (2014) estimated that males carrying biallelic *MUTYH* mutations had a 75.4% risk of developing CRC by age 70, and females had a 71.7% risk (Win *et al* 2014).

MAP may also have extra-colonic manifestations, although these are generally not part of the characteristic phenotype (Poulsen and Bisgaard 2008). The lesions which have been reported include duodenal adenomas and carcinoma, fundic gland polyps, stomach cancer, CHRPE, osteomas and breast cancer (reviewed in Venesio *et al* 2012; reviewed in Poulsen and Bisgaard 2008). The incidence of extra-intestinal malignancies is almost double that of the general population, with a significant increase in ovarian, bladder and skin cancers (reviewed in Venesio *et al* 2012).

By 2013, >300 *MUTYH* variants among MAP patients and/ or controls had been described (Ruggieri *et al* 2013). The mutations observed in *MUTYH* vary according to the ethnic group studied, suggesting population specific ancestral variants (Dolwani *et al* 2007; Sieber *et al* 2003). In Caucasian populations, P.Tyr165Cys and p.Gly382Asp are the common mutations (Jones *et al* 2002; Sampson *et al*

2003; Enholm *et al* 2003; Sieber *et al* 2003). These account for 73% of all mutations reported and have been described in Swiss (Russell *et al* 2006), Italian (Gismondi *et al* 2004; Venesio *et al* 2004), French (Küry *et al* 2007), Swedish (Kanter-Smoler *et al* 2006; Zhou *et al* 2005), Canadian (Croitoru *et al* 2007), Australian (Kairupan *et al* 2005), Portuguese (Isidro *et al* 2004), Czech (Šulová *et al* 2007), British, American and Dutch populations (reviewed in Cheadle and Sampson 2007). There is evidence for strong founder effects for these mutations: it is suggested that they derive from ancestors who lived between 5-8000 years and 6-9000 years BC respectively (Aretz *et al* 2014).

Numerous other pathogenic variants in *MUTYH* have been reported, including c.1395delGGA in Italians (Gismondi *et al* 2004), p.Glu466* and p.TyrY90* in Asians (Sampson *et al* 2003; Jones *et al* 2002) and p.Arg231Cys in the Japanese (Miyaki *et al* 2005).

1.5.1.3 Polymerase Proofreading Associated Polyposis (PPAP)

PPAP is a relatively recently defined clinical entity. *POLE* and *POLD1* code for DNA polymerases with exonuclease (proofreading) activity. Mutations in these genes are thought to cause a defect in correcting mispaired bases inserted during DNA replication (Palles *et al* 2013). In 2012, Palles *et al* undertook whole-genome sequencing of probands who had at least 10 colorectal adenomas by age 60, who had previously had known Mendelian cancer syndromes excluded in a clinical diagnostic setting. They also sequenced several affected relatives (Palles *et al* 2013). The group found that a genetic variant, *POLE* p.Leu242Val, was associated with multiple colorectal adenomas and carcinoma (Palles *et al* 2013). The trait showed dominant inheritance, with high penetrance (Palles *et al* 2013). Another variant, *POLD1* p.Ser478Asn, predisposed to colorectal tumours, endometrial cancer, and possibly brain tumours (Palles *et al* 2013).

In 2014, Valle *et al* sought to determine the prevalence of these mutations in 858 patients with unexplained familial/ early-onset CRC or polyposis. They didn't identify either mutation in any of their CRC cases. However, the *POLE* p.Leu424Val mutation was found in a polyposis family, in which case it had occurred as a *de novo* mutation in the proband. This accounted for 0.52% of the polyposis cases.

The group also reported a novel variant, *POLD1* c.1421T>C (p.Leu474Pro) in a MMR-proficient family. This mutation occurs in the proofreading domain of a DNA polymerase and was predicted to be pathogenic (Valle *et al* 2014).

Valle's group sought to further characterise the phenotypic spectrum of patients carrying germline *POLE/ POLD1* mutations (Bellido *et al* 2016). They sequenced the entire exonuclease domains of *POLE/ POLD1* in 544 CRC cases from 529 families, including those from Valle's original paper (Valle *et al* 2014). Although no additional *POLE* mutations were identified, 4 of 6 novel/ rare nonsynonymous *POLD1* variants detected were believed to be pathogenic: p.Asp316His, p.Asp316Gly, p.Arg409Trp and p.Leu474Pro. The group reviewed the phenotypic data from all 69 carriers of *POLE/ POLD1* mutations that had been reported to date. They observed that the associated phenotype was characterised by attenuated/ oligo- adenomatous polyposis, with >80% of *POLE* and >60% of *POLD1* mutation carriers being diagnosed with >/= 2 adenomas, with an average of 19 lesions. CRC was diagnosed in 60-64% of carriers, and brain tumours in 5.8%. Gastroduodenal (mostly duodenal) adenomas were identified in 57.1% of carriers who underwent gastroduodenoscopies. For patients harbouring *POLD1* mutations, the phenotypic spectrum was extended to include endometrial tumours (57.1% of carriers) and breast tumours (14.3% of carriers).

1.5.1.4 *NTHL1*-Associated Polyposis

A further gene involved in the pathogenesis of colorectal neoplasia is *NTHL1.* In 2015 Weren *et al* carried out whole exome sequencing on 51 patients with multiple colorectal adenomas +/- CRC, who had tested negative for *APC* and *MUTYH* mutations (Weren *et al* 2015). The group found that 7 individuals were homozygous for a *NTHL1* nonsense mutation, c.268C>T, which triggers nonsense-mediated decay (NMD). The patients harbouring the mutations all had multiple colorectal adenomas, ranging from 8-50, and 4 also had multiple CRCs. All 3 affected women developed complex endometrial hyperplasia or endometrial cancer.

NTHL1 is a base excision repair gene, and homozygous mutations cause an increase in C:G>T:A changes in genes such as *APC, p53, KRAS* and *PI3K* (Weren *et al* 2015).

1.5.1.5 *MSH3*-Associated Polyposis

The most recently identified polyposis syndrome is *MSH3*-Associated Polyposis (Adam *et al* 2016). Adam's group performed whole exome sequencing on germline DNA extracted from 102 unrelated individuals with unexplained adenomatous polyposis. They found two different individuals with different compound heterozygous mutations in the mismatch repair gene, *MSH3.* Both index persons had an affected sibling carrying the same mutations. The mutations were associated with tumours which displayed Elevated Microsatellite Alterations At Selected Tetranucleotide Repeats (EMAST), a type of microsatellite instability.

The phenotypic spectrum in *MSH3* mutation carriers was reported to include colorectal and duodenal adenomas, CRC, gastric cancer and an early onset astrocytoma (Adam *et al* 2016).

1.5.2 Hyperplastic Polyps (HPPs)

HPPs are a frequent finding, seen in between 1% and 73% of autopsies (Paspatis *et al* 2001; Williams *et al* 1982; reviewed in Correa *et al* 1977). In asymptomatic individuals undergoing colonoscopy, HPPs are observed in 21% - 34% of cases (Forsberg *et al* 2012; DiSario *et al* 1991). They are more common in males than females, and their prevalence increases with age (Williams *et al* 1982; Williams *et al* 1980; Correa *et al* 1977).

Until approximately 1990, hyperplastic (or 'metaplastic') polyps were regarded as a homogeneous group of tumours with no malignant potential (Rosty *et al* 2013a). Since that time, it has been increasingly recognised that hyperplastic lesions are not a single entity – they differ in their morphology and their clinical significance, in particular their risk for progressing to carcinoma. In 2010, the World Health Organisation published a classification system which subdivides hyperplastic lesions into 3 groups based upon their microscopic appearance: Hyperplastic Polyps (HPPs), Sessile Serrated Adenomas/ Polyps with or without cytological dysplasia (SSA) and Traditional Serrated Adenomas (TSA) (reviewed in Rosty *et al* 2013a; reviewed in Leggett and Whitehall 2010).

It is now known that hyperplastic lesions may be the precursors to CRC developing along the serrated pathway of carcinogenesis, which accounts for approximately

10% of CRCs (Yamane *et al* 2014). In contrast to the CRC which follow the traditional 'adenoma carcinoma' pathway, tumours arising from the serrated pathway tend not to display CIN, but instead exhibit MAPK pathway activation, through *BRAF* mutations, and they commonly develop CIMP (Yamane *et al* 2014; Rosty *et al* 2013a; Leggett and Whitehall 2010).

1.5.2.1 Serrated Polyposis Syndrome

Although most HPPs are sporadic lesions, there is a condition in which patients develop multiple and/ or large lesions. One of the first descriptions was in 1980 by Williams *et al* (Williams *et al* 1980). They observed 7 patients, with a mean age of 37.4 years, who each had at least 50 lesions throughout their large bowel. At that time, the authors concluded that 'it is impossible to deduce whether or not 'metaplastic polyposis' is a distinct entity. There is no good evidence that it is familial in this small series, but the appearance of numerous metaplastic polyps of an unusually large size and configuration, predominantly in young males, might suggest a specific 'disease''.

As increasing evidence came to light that there seemed to be a syndrome in which patients developed numerous hyperplastic lesions throughout their large bowel, it was named 'hyperplastic polyposis syndrome' (HPS). In 2000, a definition of HPS was proposed by Jass and Burt in the World Health Organisation classification of tumours (Jass and Burt 2000). This definition was modified in 2010, and the disease was officially renamed Serrated Polyposis Syndrome (SPS). It appears in the 2010 World Health Organisation classification of tumours of the digestive system (Snover DC *et al* 2010). In order to meet the diagnostic criteria for SPS, patients must fulfil at least one of the following criteria:

1. At least 5 serrated (hyperplastic) polyps proximal to the sigmoid colon, 2 of which are >10mm diameter
2. Any number of serrated polyps proximal to the sigmoid colon in an individual who has a first degree relative with serrated polyposis
3. More than 20 serrated polyps of any size distributed throughout the colon

(reviewed in Rosty *et al* 2013a; reviewed in Guarinos *et al* 2012; reviewed in Leggett and Whitehall 2010; Snover DC *et al* 2010)

It was thought that SPS affected 1 in 3000 asymptomatic individuals between the ages of 55 and 64 years (reviewed in Rosty *et al* 2013b). However, most studies report a broad age distribution of the disease (17 to 85 years) with a mean age of diagnosis of 47.7-56 years, so it is possible that the prevalence of SPS in the general asymptomatic population is higher than 1/3000 (Rosty *et al* 2013b). Recent data suggest that the prevalence may be as high as 1/151 patients who have a colonoscopy following a positive faecal occult blood test (reviewed in Rosty *et al* 2013a).

SPS shows no sex predilection, and the mean age of diagnosis is 55 years (Guarinos *et al* 2012; Kalady *et al* 2011). As well as hyperplastic lesions, up to 85% of patients also have conventional adenomas present in the bowel (Rosty *et al* 2013a; Rosty *et al* 2012). Patients with SPS have an increased risk of developing CRC, which generally occurs between 50 and 60 years of age (reviewed in Guarinos *et al* 2012). Malignancy is associated with a larger number of polyps, the presence of dysplasia (reviewed in Guarinos *et al* 2012; Yeoman *et al* 2007) and the presence of conventional adenomas in addition to HPPs (Rosty *et al* 2013b). The incidence of CRC in SPS patients varies from 14% to 58% (reviewed in Rosty *et al* 2013b; Yeoman *et al* 2007; Hyman *et al* 2004; Lage *et al* 2004) and the incidence is greater in females than in males, with a ratio of 2.4:1 (Rosty *et al* 213b). When carcinoma develops in an SPS patient, it is likely to have a proximal location: 64% of CRC are identified proximal to the descending colon (Rosty *et al* 2013b). Interestingly, a large proportion of CRCs seen in patients with SPS do not develop through the 'serrated pathway of carcinogenesis' driven by *BRAF* mutation (Rosty *et al* 2013b). The tumours show various molecular changes, including those more likely to be associated with the traditional adenoma-carcinoma pathway, for example β-catenin activation and/ or overexpression of p53 (Rosty *et al* 2013b).

SPS is thought to be a genetic disease, but the mode of inheritance is unclear (reviewed in Guarinos *et al* 2012). There are papers which report germline mutations in the Wnt inhibitor *RNF43*, an E3 ubiquitin ligase, in individuals and families with features of SPS (Yan *et al* 2017; Taupin *et al* 2015; Gala *et al* 2014). However, Buchanan *et al* note that mutations in *RNF43* may account for only a small proportion of SPS, and that additional genetic risk factors are yet to be identified (Buchanan *et al* 2017).

1.5.3 Hamartomatous Polyps

Hamartomas are overgrowths of the tissue which is native to the site of the lesion. Polyposis syndromes which are characterised by hamartomas include Peutz-Jegher's Syndrome (PJS), Juvenile Polyposis Syndrome (JPS) and Cowden's Disease.

1.5.3.1 Peutz-Jeghers Syndrome (PJS)

In 1921, a Dutch physician, Dr. Peutz, first described the combination of gastrointestinal polyps and mucocutaneous pigmentation (Peutz 1921). In 1949, Dr. Jeghers published an article describing 10 patients who had a combination of pigmentation of the oral mucosa/ lips/ digits, and intestinal polyps (Jeghers *et al* 1949). The observations made by Peutz and Jeghers led to the definition of an AD syndrome characterised by gastrointestinal polyposis and mucocutaneous pigmentation, now known as Peutz-Jeghers Syndrome (PJS) (Westerman *et al* 1999).

PJS is inherited in an AD manner and has variable penetrance. It is a rare condition, with a prevalence of approximately 1 in 200 000 (reviewed in Omundsen and Lam 2012). The disease is characterised by hamartomatous polyps throughout the gastrointestinal tract and mucocutaneous pigmentation. Patients present at a median age of 11 years, and this is often as a result of a complication of their GI polyps, for example intussusception, small bowel obstruction, rectal bleeding or volvulus (reviewed in Omundsen and Lam 2012).

Approximately 50% of cases of PJS are caused by germline mutations in the nuclear serine threonine kinase gene *LKB1/ STK11* (Jenne *et al* 1998; reviewed in Omundsen and Lam 2012). This gene regulates cell polarisation, growth and metabolism. Most mutations are small insertions or deletions, resulting in a truncated protein with no kinase activity (reviewed in Omundsen and Lam 2012).

Patients with PJS are at increased risk of developing cancer, both at gastrointestinal and extra-intestinal sites. The most common tumours are CRCs, but there is also an increased risk of other gastrointestinal (GI) carcinomas (oesophageal, gastric, small bowel and pancreas), breast cancer, cervical cancer and sex cord tumours (Hizawa *et al* 1993; reviewed in Omundsen and Lam 2012).

1.5.3.2 Familial Juvenile Polyposis Syndrome

In 1914, Hertz described four family members who had rectal polyps in childhood, the youngest being only 8-years-old (Hertz 1914). That report is regarded as the first instance of juvenile polyposis in the medical literature (Calva and Howe 2009). Juvenile polyps are common hamartomatous lesions which occur in the large bowel. They are usually solitary and sporadic. If multiple juvenile polyps are present, the patient may have Juvenile Polyposis Syndrome (JPS). JPS is an AD disease with variable penetrance and which has an incidence of 1 per 100 000 births (reviewed in Omundsen and Lam 2012). Patients can present in infancy with GI bleeding, intussusception, rectal prolapse or a protein losing enteropathy. Around 15% will have an associated congenital birth defect, such as gut malrotation, cardiac and cranial abnormalities, cleft palate, polydactyly or genitourinary defects. If presentation is as an adult, the patient is likely to suffer from GI bleeding (reviewed in Omunsden and Lam 2012).

For a diagnosis of JPS, patients must fulfil one of the following criteria:

1. More than 5 juvenile polyps of the colon or rectum
2. Juvenile polyps in other parts of the GI tract
3. Any number of juvenile polyps and a positive family history

(reviewed in Omundsen and Lam 2012)

Germline mutations in *SMAD4* and *BMPR1A* are seen in JPS, and it is suggested that *ENG* mutations may also have a role (Sweet *et al* 2005), although this is not certain (Howe *et al* 2007). All of these genes are involved in transforming growth factor β (TGFβ) signalling. The TGFβ family of cytokines are growth inhibitors, and loss of sensitivity to these factors promotes tumourigenesis (reviewed in Fleming *et al* 2013). As such, patients with JPS are at increased risk of developing CRC (Rozen and Baratz 1982; Järvinen and Franssila 1984; Giardiello *et al* 1991). In additional to colorectal malignancies, patients are also at risk of developing gastric and duodenal cancer (reviewed in Omundsen and Lam 2012).

1.5.3.3 Cowden's Syndrome

In 1963 Lloyd and Dennis reported a 20-year old female with multiple pathologies, including multiple thyroid adenomas, extensive fibrocystic change of both breasts

and ‘space occupying lesions in the liver and bone’ (Lloyd and Dennis, 1963). At that time, it was noted that ‘whether this case represents a new familial disease … has not been established’ (Lloyd and Dennis, 1963), although they named the syndrome ‘Cowden’s Disease’. It is now established that Cowden’s Syndrome is an AD disease, affecting approximately 1 in 200 000 births. It is commonly diagnosed in the second decade of life, but age of onset may vary from 4 to 75 years (reviewed in Lam-Himlin *et al* 2014). Patients develop multiple hamartomas in multiple organ systems. Lesions can be found in the skin, GI tract, breast, thyroid gland and central nervous system (reviewed in Omunsden and Lam 2012; Hanssen and Fryns 1995).

80% of patients have a germline mutation in the *PTEN* gene (reviewed in Omunsden and Lam 2012). *PTEN* is a tumour suppressor: its product is a phosphatase which negatively regulates the phosphatidylinositol 3-kinase-AKT and mammalian target of rapamycin (mTOR) signalling pathways, which are involved in cell growth and proliferation, cell cycle progression and apoptosis (reviewed in Lam-Himlin *et al* 2014). *PTEN* mutations are involved in the pathogenesis of several carcinomas, including breast, endometrial, thyroid, large bowel and kidney (reviewed in Lam-Himlin *et al* 2014). As such, patients with Cowden’s Syndrome are at increased risk of these malignancies (Hanssen and Fryns 1995).

1.5.3.4 Hereditary Mixed Polyposis Syndrome (HMPS)

Hereditary Mixed Polyposis Syndrome (HMPS) is a relatively recently defined entity. In 1971, Kaschula described an 11-year-old girl who had profuse diarrhoea mixed with blood and mucus. She was found to have polyps throughout her large bowel, and the polyps had both adenomatous and juvenile morphologies (Kaschula 1971). Over a decade later, in 1987, the term ‘mixed familial polyposis syndromes’ was used as the title of a report by Sarles *et al* (Sarles *et al* 1987). This article described 3 patients, including a father and son, who all had multiple polyps of different histopathological types.

In HMPS, patients develop multiple polyps with mixed morphologies. This is an AD disease, and patients may have adenomas, hyperplastic polyps and hamartomatous polyps. There is a high risk of developing CRC (Jaeger *et al* 2012). The disease is caused by a duplication spanning part of the *SCG5* gene and a region upstream of

the *GREM1* locus. This duplication causes increased expression of *GREM1,* which acts as a bone morphogenetic protein (BMP) antagonist. The subsequent reduction in BMP signalling is thought to play a role in tumourigenesis (Jaeger *et al* 2012).

1.6 Clinical Management of Patients with Colorectal Polyposis

This thesis is focused on patients with multiple colorectal adenomas. In view of the possible genetic diagnoses which may underlie such a phenotype it has been common clinical practice for patients with >10 colorectal adenomas to be referred to a regional genetics centre for genetic counselling and for consideration of diagnostic analysis of the *APC* and/ or *MUTYH* genes. The diagnostic testing carried out will depend upon the individual's phenotype and their family history.

Up to 90% of patients with a phenotype of typical FAP have a pathogenic *APC* germline mutation identified through sequencing of coding exons and deletion/ duplication analysis via multiplex ligation-dependent probe amplification (MLPA) (Spier *et al* 2012). Of those with a phenotype of AFAP, *APC* or biallelic *MUTYH* germline mutations are detected in only 20-50% of cases (Spier *et al* 2012).

1.6.1 Why Might Some Genetic Variants Be Missed?

Since 1991, when *APC* was recognised as the causative gene of FAP (Groden *et al* 1991; Joslyn *et al* 1991; Kinzler *et al* 1991; Nishisho *et al* 1991), several screening and diagnostic strategies have been developed to identify pathogenic *APC* mutations (Scott *et al* 2001) in patients with multiple colorectal adenomas. These have included denaturing gradient gel electrophoresis analysis (DGGE) (Scott *et al* 2001; Olschwang *et al* 1993), ribonuclease protection analysis (Miyoshi *et al* 1992a), single strand conformation polymorphism analysis (SSCP) (Cottrell *et al* 1992; Groden *et al* 1993), heteroduplex analysis (HA) (Cottrell *et al* 1992) and the protein truncation test (PTT)/ in vitro synthesised protein assay (IVSP) (Powell *et al* 1993). The characterisation of a genetic mutation identified through screening requires DNA sequencing. Sanger sequencing has been the gold standard of sequencing for several decades, and until very recently was the main approach used for the molecular diagnosis of colorectal polyposis. In the last few years, next-

generation sequencing (NGS) technologies have progressively replaced Sanger sequencing.

Genetic diagnostic protocols used in patients with polyposis typically include sequencing of *APC* and/ or *MUTYH* and dosage analysis of the genes using a technique such as MLPA. These approaches could miss pathogenic variants located in promoter regions, deep within introns (Spier *et al* 2012) or in untranslated regions (UTRs), which may have effects on gene expression mediated through effects on transcription, mRNA splicing or mRNA stability. Similarly, diagnostic protocols may miss low frequency variants in patients with somatic mosaicism. Protocols used in the National Health Service (NHS) in Wales would also not identify epigenetic phenomenon such as promoter methylation, they would not detect mutations in genes which are established but rare causes of polyposis (e.g. *POLE*/ *POLD1*) and they would not identify novel polyposis genes.

1.6.1.2 Promoter Variants and Allelic Imbalance (AI)

The *APC* gene has two promoter regions, 1A and 1B (reviewed in Rohlin *et al* 2011). The major transcript is initiated by the major promoter, 1A (reviewed in Charames *et al* 2008). It is possible that genetic variants occurring in these promoters could lead to reduced gene expression, therefore predisposing to tumour formation. Such variants might not be identified through standard genetic diagnostics as the promoter is not typically included in diagnostic sequencing protocols.

There are several reports of *APC* promoter mutations in the literature. However, these typically describe deletions, which would be detected through diagnostic MLPA (Yamaguchi *et al* 2016; Pavicic *et al* 2014; Rohlin *et al* 2011; Charames *et al* 2008). At the time of writing, there is a paucity of literature describing *APC* promoter point mutations or methylation as a cause of colorectal polyposis.

Allelic imbalance (AI) refers to a situation in which the two alleles of a given gene are expressed at different levels in a given cell (Wagner *et al* 2010). It can occur due to epigenetic inactivation of one of the alleles, or because of genetic variation in regulatory regions (Wagner *et al* 2010). AI can involve complete inactivation of one allele, for example in parent-of-origin imprinting, when a specific allele at a given

locus is silenced through epigenetic mechanisms depending on whether it was inherited from the mother or father (Wagner *et al* 2010). Allele expression can also be partially reduced, which can occur when different alleles have differing affinities for transcription factors (Wagner *et al* 2010), or through *cis*- acting genetic variants, for example in promoters (Wagner *et al* 2010).

As well as being important in normal phenotypic variation, AI can also contribute to tumourigenesis. There is some evidence that *BRCA1* and *BRCA2* AI plays a role in the pathogenesis of ovarian and breast cancer (Chen *et al* 2008; Shen *et al* 2011). Whilst only a small number of studies have considered *APC* AI in the context of colorectal neoplasia, those which have been performed have found it may make an important contribution. As early as 1993, Powell *et al* (Powell *et al* 1993) used an allele-specific expression assay to show that 3/11 *APC* NMI patients with clinical FAP had significantly reduced expression of one *APC* allele. In 1999 Laken *et al* (Laken *et al* 1999) used monoallelic mutation analysis (MAMA) to reveal that 7/9 *APC* NMI patients had reduced/ no expression from one of their *APC* alleles. More recently Yan *et al* (2002b) identified a patient with colorectal tumours who was known to have reduced levels of the APC protein. The group quantified the relative levels of mRNA transcripts from each *APC* allele using Digital-SNP. They found that gDNA yielded the expected 50% allelic ratio, but that cDNA from lymphoblastoid cells showed a skewed distribution, with a ratio of approximately 66% (Yan *et al* 2002b). Linkage analysis showed that the allele whose mRNA was expressed in lower amounts was the one linked to disease (Yan *et al* 2002b). Further work confirmed that the skewed allelic ratio was also present in 4 affected family members, but that the ratio was normal in 24 unrelated unaffected individuals. The group continued to investigate expression levels of *APC*: in four patients with clinical FAP who had no abnormalities with the *in vitro* synthesised protein assay (IVSP) or allele sequencing, one was found to have an abnormal 71% allelic ratio in cDNA. When tumours from the patients with AI were studied, 30/38 had LOH of *APC,* and in 29 of these cases, it was the normal allele which had been lost. Yan *et al* concluded that an allele which causes a decrease in transcript levels can result in a predisposition to severe disease, but that there needs to be a second hit to the normal allele for a disease to manifest (Yan *et al* 2002b). Interestingly the cause of the decreased expression was not determined: the sequences of the coding regions, promoter and 3'UTR were normal, so the group assumed that the pathogenic variant must lie within an intron or upstream of the gene (Yan *et al* 2002b). These early findings regarding *APC* AI have been supported by

Castellsagué *et al* (2010). Of 23 *APC*/ *MUTYH* NMI polyposis families who were heterozygous for rs2229992, 2 were shown to harbour *APC* AI. The AI in one family was suggested to result from promoter variants (Castellsagué *et al* 2010).

1.6.1.3 Intronic Variants

Intronic mutations may affect RNA splicing and introns are not screened as part of most routine genetic diagnostic protocols. In 2000 Su *et al* sought to identify novel intragenic rearrangements of *APC* in patients with a clinical phenotype of FAP or AFAP. They found four germline *APC* mutations, one of which was a deletion 27-1627bp downstream of exon 14, which was replaced with a novel sequence of about 180bp. The deletion was completely within intron 14, but it affected the splicing of exon 14 (Su *et al* 2000). In 2010 Tuohy *et al* used Southern Blot analysis of the *APC* gene to identify a 1.4kb deletion within intron 14 in a family with AFAP. Subsequent PCR amplification from exon 13 to exon 15 of cDNA showed that the intronic deletion resulted in abnormal splicing, and that exon 14 was deleted. This caused a frameshift and protein truncation at codon 673 of the normal reading frame resulting in a truncated product that lacked all of the β-catenin, microtubule and EB-1 binding domains (Tuohy *et al* 2010).

In 2012 transcript analysis in a sample of 125 mutation negative patients with colorectal adenomatous polyposis found that 8% had a reproducible aberrant transcript pattern, suggesting an intronic mutation at a genomic level (Spier *et al* 2012). 80% of these were found to have transcript insertions between two exons originating from exonised sequences deep within the corresponding intron (Spier *et al* 2012). All pseudoexons were predicted to result in out-of-frame transcripts with premature stop codons (Spier *et al* 2012). In those patients who had insertions, the underlying genomic mutations were identified: they comprised 3 different heterozygous point mutations (c.532-941G>A, c.1408+731C>T, c.1408+735A>T) which activated cryptic splice sites. A pre-existing complementary cryptic splice site was predicted at the other end of the insertion (Spier *et al* 2012).

1.6.1.4 Untranslated Region (UTR) Variants

A pre-RNA molecule undergoes several steps of processing before it becomes a functional mRNA molecule (Mignone *et al* 2002). Mature mRNA consists of a 5'

untranslated region (5'UTR), a coding region, and a 3' UTR (Mignone *et al* 2002). UTRs, particularly 3'UTRs, have multiple roles in the post-transcriptional regulation of gene expression, including effects on mRNA transport out of the nucleus, translation efficiency, subcellular localisation and mRNA stability (reviewed in Mignone *et al* 2002). In addition to normal physiological intracellular effects, there is increasing evidence that UTR variants can be involved in disease, for example an expanded number of trinucleotide repeats in the 3'UTR of the *DMPK* gene is thought to play a role in the pathogenesis of in myotonic dystrophy (Conne *et al* 2000), and a somatic 5'UTR variant has been reported to reduce translation efficiency of *BRCA1* in a highly aggressive sporadic breast cancer (Signori *et al* 2001).

In the context of colorectal neoplasia, Wilding *et al* have shown that in microsatellite unstable cancer, deregulation of mRNA stability due to mutations in regulatory 3'UTR sequences can lead to a marked difference in gene expression profiles when compared to microsatellite stable tumours (Wilding *et al* 2010).

1.6.1.5 Mosaicism

A mosaic is an individual who has at least two genetically different cell lines despite developing from a single zygote. Mosaic mutations may be missed with standard mutation diagnostic techniques, for example if they occur at a low frequency within the individual or if they do not occur in the part of the body which is being analysed.

It is reported that somatic mosaicism can occur in 10-20% of sporadic cases of FAP (reviewed in Rohlin *et al* 2009; Hes *et al* 2008; Aretz *et al* 2007). The timing at which an *APC* mutation occurs will have an important bearing on the patient's phenotype: if it arises in a single colonic epithelial stem cell, the only consequence will be adenomatous polyps in the segment of the colon that becomes populated with descendants of the stem cell (Tuohy and Burt 2008). At the other extreme, if the mutation occurs early in embryogenesis, it may be found in all three germ cell layers. This would result in multiple clinical manifestations seen throughout multiple organ systems, potentially including mutations in reproductive cells, which could then be passed on to future generations (Tuohy and Burt 2008).

Depending on the frequency of the mutation, it is likely that a significant proportion will be missed using current diagnostic sequencing protocols and DNA extracted

from whole blood. Rohlin *et al* (2009) carried out a study to evaluate the different mutation screening/ diagnostic techniques in terms of their sensitivity in detecting mosaicism. They looked at Sanger sequencing, single-strand conformation polymorphism (SSCP)/ heteroduplex analysis (HA), the protein truncation test (PTT), denaturing high performance liquid chromatography (DHPLC) and massively parallel sequencing. A total of 9 mutations were addressed – 8 in *APC* and 1 in *NF2.* The group constructed 7 artificial mosaics in *APC* through serial dilutions of DNA, with a non-mosaic heterozygous mutation being defined as 100%. The two remaining samples were from naturally occurring mosaics. All of the dilutions of all of the mutations were analysed with SSCP, DHPLC and Sanger sequencing. Three were included in the PTT assay (these were mutations in exon 18, which is readily screened by PTT). Only four artificial mosaics, at various concentrations, and both natural mosaics underwent massively parallel sequencing, due to cost limitations. The group found that SSCP and DHPLC were able to detect mutant alleles at frequencies between 5% and 25%, whereas Sanger sequencing required frequencies between 15% and 50% for detection. The mutations included in the PTT assay were detected at frequencies between 10 and 100%. The Genome Sequencer FLX was used for massively parallel sequencing, and this achieved coverage between 648 and 8313 reads. Mutation frequencies as low as 1% could be detected, but this required a high coverage (Rohlin *et al* 2009).

The results from this study showed that Sanger sequencing, which has been commonly used in a diagnostic setting, was the least sensitive method at detecting mosaics. Dependent upon the type of mutation being analysed, this technique may require mutation frequencies as high as 50% in order for them to be detected (Rohlin *et al* 2009).

Genetic testing in the diagnostic setting is usually carried out on DNA extracted from whole blood. For some patients, a mosaic mutation may only be present in colonic tissue so will not be detectable through testing blood-derived DNA (Jansen *et al* 2017).

1.6.1.6 The Involvement of Additional Genes

Standard diagnostic protocols carried out in the context of adenomatous colorectal polyposis typically examine the *APC* and *MUTYH* genes, although an increasing

number of laboratories are screening DNA for the recurrent mutations in *POLE* and *POLD1.* It is only in recent years that the pathogenic effects of mutant POLE, POLD1, NTHL1 and MSH3 have been discovered. It is therefore feasible that there may be further genes involved in the development of heritable colorectal polyposis, which are yet to be identified.

Furthermore, mutations in genes such as *APC, MUTYH, POLE, POLD1*, *NTHL1* and *MSH3* are all highly penetrant. It may be the case that some patients with unexplained colorectal polyposis have a phenotype which results from the complex interplay of several low/ moderate penetrance genetic variants.

1.7 Summary

Colorectal carcinoma (CRC) is the fourth most common cancer in the United Kingdom (UK) and the second most common cause of cancer death (Cancer Research UK 2015). Most CRCs occur sporadically, but in approximately a third of patients, hereditary factors are important (reviewed in Mishra and Hall 2012; Burt 2007). Some patients with an inherited predisposition to CRC will be diagnosed with a 'genetic polyposis syndrome' such as FAP, MAP, PPAP, *NTHL1*-Associated Polyposis, *MSH3*-Associated Polyposis or a hamartomatous polyposis syndrome. It is important to identify these patients, and to define the mutations causing their polyposis, so that the individuals and their relatives can be managed appropriately.

1.8 Aims of Thesis

The genetic bases for several 'polyposis syndromes' are known and have been described above. However, there are numerous patients with a 'polyposis phenotype' who have had no genetic mutation identified (NMI) through standard genetic diagnostic protocols.

The aim of this thesis is to:

1. Identify novel genetic variants/ molecular mechanisms responsible for colorectal polyposis using a cohort of polyposis patients who have had no mutation identified during clinical genetic diagnostic testing. This will involve:

a. searching for *APC* and *MUTYH* variants which are outside of the open reading frame (ORF)
b. searching for mosaic *APC* mutations
c. searching for novel variants in the other known or candidate 'polyposis genes'
d. searching for new polyposis genes
e. characterising novel variants using genetic and functional approaches, to provide additional evidence for or against their clinical significance

Chapter 2 Materials and Methods

2.1 Materials, Equipment and their Suppliers

2-Propanol	Sigma Aldrich
2X iproof HF Mastermix	BioRad
25cm^2 flask	ThermoFisher
96-well Assay plate	Fisher Scientific
96-well PCR plates	4titude
96-well PCR seals	4titude
ABI Prism Genetic Analyser	Applied Biosystems
Acetic Acid	Sigma Aldrich
Agarose	ThermoFisher
Agarose plates	Millipore
Ampligase	Illumina
Ampligase DNA Ligase Buffer	Illumina
Anti-β-Actin Antibody (A5411)	Sigma
Anti-Mouse Antibody (NXA931)	GE Healthcare
Anti-Myc-Tag Antibody (9B11)	Cell Signalling Technologies
β-mercaptoethanol	Stratagene
BCA	ThermoScientific
Big Dye Terminator Reaction Mix v1.1	ThermoFisher
BioDoc-IT Imaging System	AnalytikJena
BSA	Promega
Cell Scraper	ThermoFisher
Centrifuge PIC017	ThermoScientific
Chemagic STAR DNA Extraction Kit	ThermoFisher
Complete protease inhibitor	Roche
Control Methylated DNA	Qiagen
Control Unmethylated DNA	Qiagen
DMEM (1X)-GlutaMax	Life Technologies
dNTPs	ThermoFisher
DTT	ThermoFisher

EDTA	Sigma Aldrich
Epitect Bisulfite Kit	Qiagen
Epitect Mastermix	Qiagen
Eppendorf Tubes	ThermoFisher
Exonuclease I/III	New England Biolab
Falcon Tubes	ThermoFisher
FBS	ThermoFisher
FS Buffer	ThermoFisher
G Storm Labtech Model GS0002M	Labtech
Gel Red	Cambridge Bioscience
Gel Tank	ThermoScientific AB0708 100V/ Clever Scientific
Gene Read FFPE Kit	Qiagen
Gene Ruler 1kb Plus Ladder	ThermoFisher
Glo-Lysis Buffer and Substrate	Promega
Glycerol	Sigma Aldrich
Haloplex Assay	Agilent
Hemo Klen Taq DNA Polymerase	New England Biolab
HiDi Formamide	ThermoFisher
HiSeq 2500	Illumina
Iblot	Invitrogen
IGEPAL CA630	Sigma Aldrich
Incubator	Millipore
Light Microscope	Both Olympus BX43 and Leica models have been used
Lipofectamine	Life Technologies
LB Agar Lennox	Invitrogen
LB Broth Base Lennox	Invitrogen
Milk Powder	Marvel
MiniPrep Kit	Qiagen
MiSeq	Illumina
MLPA Kit P043-B1	MRC Holland
MLPA Kit P378-A2	MRC Holland
MMG	Microzone
NaCl Solution	Sigma Aldrich
Nextera Rapid Capture Exome v1.2	Illumina

Novex Sharp Pre-Stained Protein Standard	Invitrogen
NuPAGE 4-12% Bis Tris Mini Gel	Invitrogen
NuPAGE MES SDS Running Buffer	Invitrogen
NuPAGE LDS Sample Buffer	Invitrogen
Orange G	Sigma Aldrich
Orange G Solution (20ml)	4 ml 0.5 M EDTA, 8 g Sucrose, pinch of Orange G, Water
Optimem	Life Technologies
PBS	Santa Cruz Biotech
PhosphoSTOP Phosphatase Inhibitor	Roche
Pipettes and Tips	Rainin
Platinum PFX PCR Kit	Invitrogen
Power Supply	BioRad 200/2.0
QBD2 Heat Block	Grant
Quantstudio 12K Real Time PCR System	Life Technologies
Qubit dsDNA BR Assay Kit	ThermoFisher
Qubit dsDNA HS Assay Kit	ThermoFisher
Qubit 2.0 Fluorometer	ThermoFisher
Qubit RNA Assay Kit	ThermoFisher
QuikChange II SDM Kit	Agilent
Random Primers	ThermoFisher
Restore Plus Western Blot Stripping Buffer	ThermoFisher
RNASin	ThermoFisher
R-Spondin	PeproTech
Shaking Incubator	Kühner
Shrimp Alkaline Phosphatase	Affymetrix
SOC Medium	Invitrogen
Sucrose	Sigma Aldrich
Superscript II	ThermoFisher
Supersignal West Pico Chemiluminescent Substrate	ThermoFisher
TAE (1L)	48.4 g Tris, 3.72 g EDTA, 17 ml Acetic Acid, Water
Taqman Gene Expression Mastermix	ThermoFisher
TBS	Fisher BioReagents
TC Hood	HeraSafe

TC Incubator (37˚C, 5% CO_2, 21% O_2)	Binder
TE Buffer	Invitrogen
Tempus Blood RNA Tube	Life Technologies
Tempus Spin RNA Isolation Reagent Kit	Applied Biosystems
Tris	Sigma Aldrich
Trypsin (0.05%)	Life Technologies
Trypan Blue (0.4%)	Life Technologies
Vortex Genie 2	Scientific Industries
Waterbath	Clifton
Wnt3a	R&D Systems
WST-1	Sigma Aldrich
XL-1 Blue Cells	Agilent
Zymoclean Gel DNA Recovery Kit	Zymo Research

2.2 Patient Information

This study is approved by the Research Ethics Committee for Wales (REC 3, study 12/WA/0071). To date, 306 patients of the 350 target have been recruited. A subset of these patients was used for the work described in this thesis.

2.2.1 Inclusion Criteria

Eligible participants included:

1) adult patients who are affected by polyposis (>/= 10 polyps)/ other phenotypes consistent with *APC* mutations who have had genetic testing to determine the genetic cause of their disease that proved negative by routine tests in NHS genetic diagnostic laboratories or

2) adult patients who are affected by polyposis who have previously had genetic testing to determine the genetic cause of polyposis of the bowel that found an uncharacterised variant of unknown pathogenicity by routine tests by NHS genetic diagnostic laboratories or

3) adult patients who have been identified as carrying a known polyposis-causing gene and who were having a screening or a surgical procedure at which polyp or other tumour or cancer tissue may be removed or

4) unaffected adult (aged 18 years or over) family members of these patients whose DNA could serve as a control to assess the association of genetic variants with polyposis

Participants were recruited from a pool of patients who had previously been referred for diagnostic genetic testing and/ or genetic counselling at the Institute of Medical Genetics, Cardiff, or at other regional clinical genetics centres. They were identified by consultant clinical geneticists or genetic counsellors.

2.2.2 Exclusion Criteria

The study exclusion criteria were as follows:
1) unaffected family members under the age of 18 years of age

2.2.3 Individuals Participating in this Study

Of the 306 patients recruited to the Genetic Mechanisms of Polyposis study, 60 affected individuals were included in the work described within this thesis. When the project started in 2013, these were the recruited patients for whom phenotypic information was available. I have included 45 patients with a predominantly adenomatous phenotype (>/= 10 adenomas); 8 patients with a mixed adenoma/ HPP phenotype; and 7 patients for whom histopathological information was scanty.

2.2.3.1 No Mutation Identified (NMI) Polyposis Patients

The 60 NMI polyposis patients included in this thesis have not had a pathogenic genetic mutation identified by standard genetic diagnostic protocols in the NHS (although the cohort does include two patients who are carriers of monoallelic *MUTYH* mutations). In the genetic diagnostic service, patients have undergone *APC* sequence analysis of all the coding exons and up to 20 bases of flanking intronic sequences. Dosage analysis for all coding exons and the promoter had been completed using the MRC-Holland MLPA kit P043-B1. All the patients have also undergone *MUTYH* analysis. This involved a real time allelic discrimination assay for the two common European *MUTYH* mutations (c.536A>G (p.Tyr179Cys) and c.1187G>A (p.Gly396Asp)) and Sanger sequencing of exons 3 and 14 for the

two common Asian *MUTYH* mutations (c.312C>G, p.Tyr104* and c.1432G>T, p.Glu480*). If this failed to identify any mutations, patients had sequence analysis of all coding exons and up to 20 bases of flanking intronic sequence. Dosage analysis for all coding exons was undertaken with the MRC-Holland MLPA kit P378-A2.

Blood samples for DNA were obtained for all 60 NMI patients. Blood samples for RNA were obtained for 45 of the NMI patients.

2.2.3.2 Control Samples

The control samples in this study included:

- 13 unaffected relatives of NMI patients
- 10 control FAP patients, with known *APC* mutations
- 5 control MAP patients, with known *MUTYH* mutations
- 37 healthy individuals recruited through a local study: Causes of Bowel Polyps: Recruitment of Healthy Controls (approved by Cardiff University School of Medicine Ethics Committee). The inclusion criteria were being over the age of 18 with no personal/ family history of colorectal polyposis/ CRC

DNA was available for all unaffected relatives, FAP controls and MAP controls. RNA was prepared from blood samples for 7 unaffected relatives, 3 FAP patients, 4 MAP patients and 37 healthy individuals.

2.2.3.4 Demographic Details and Clinical Phenotypes

The demographic details and clinical phenotypes of the participants are in Tables 2.1 (patients and their unaffected relatives) and 2.2 (further controls). The age of the participants refers to the age at which they were recruited to the study. This was deemed the most unambiguous way to describe age, as it was often unclear at what age the patients had initially developed polyposis, and the DNA samples used throughout this project had frequently been obtained at multiple historical time points. The phenotypic information was obtained from the patient's genetic files/

hospital records, if they had been recruited from Wales, or was supplied by referral centre if they had been recruited from another region.

Unique Identifier	**Age at Recruitment and Sex**	**Clinical Phenotype**	**RNA**	**Relative(s)**
Halo04	30-year-old female	>100 adenomas in 30s	Yes	Halo02 (mother): unaffected, negative colonoscopy
Halo05	46-year-old female	Aged 39 had multiple adenomas (at least 19): one contained pT1 CRC	No	
Halo06	77-year-old male	11 polyps: 10 adenomas, 1 HPP	Yes	
Halo07	70-year-old male	17 adenomas	Yes	
Halo08	70-year-old female	629 adenomas	Yes	S_Halo08 (son) 42-year-old male, over 100 polyps. He has not been included in the NMI patient total as he was recruited at a late stage of the project, so DNA was not available for most of the study protocols

Halo13	73-year-old male	>10 adenomas	Yes	
Halo14	33-year-old male	Subtotal colectomy and ileorectal anastomosis abroad for adenomatous polyposis. Has had a further 40 polyps removed from rectal stump ?morphology	Yes	Halo09 (sister): unaffected, negative colonoscopy
Halo15	84-year-old male	12 adenomas	No	
Halo17	66-year-old male	25 mixed polyps: adenomas and hyperplastic lesions (exact numbers of each is not clear). 2 CRC aged 63 (pT2, pT3)	Yes	
Halo18	72-year-old male	Referred from screening programme. 27 polyps: adenomas with 4 HPPs	No	
Halo19	62-year-old female	>10 adenomas and CRC (no further information available)	Yes	
Halo20	71-year-old male	Referred from screening programme. 12 adenomas, plus additional hyperplastic lesions (?number and type)	Yes	
Halo22	73-year-old female	Referred from screening programme. Aged 72 had 40+ adenomas and 2 CRC (?stage)	No	
Halo23	73-year-old female	>20 adenomas	Yes	
Halo24	68-year-old male	>10 adenomas (?19)	Yes	
Halo25	66-year-old male	12 adenomas	Yes	

Halo26	85-year-old female	21 adenomas	Yes	
Halo27	67-year-old female	17 adenomas	Yes	
Halo28	61-year-old male	1990: sigmoid colectomy for a probable diverticular perforation. Multiple 'benign metaplastic polyps' were noted ?number. 2012: subtotal colectomy for >30 polyps. The majority of the polyps were TA; 2 were serrated adenomas with LG dysplasia	No	
Halo29	80-year-old male	17 polyps: at least 9 confirmed adenomas and 7 HPPs	Yes	
Halo30	73-year-old male	20 polyps: at least 13 adenomas	Yes	
Halo31	70-year-old male	12 TA with LGD, 3 HPP	Yes	
Halo32	66-year-old female	pT2 CRC + 12 adenomas	Yes	
Halo33	36-year-old female	Heterozygous for a non-pathogenic *APC* variant: c.6363_6365dupTGC. pT3N1M1 CRC, 4 adenomas and the patient notes describe "multiple polyps in the sigmoid, smaller numbers of polyps in the proximal left colon, a couple of tiny polyps in the transverse colon and a number of small polyps in the right colon" but there is no more detail	No	

Halo34	69-year-old male	11 adenomas plus 6 polyps of unknown morphology	Yes	
Halo35	68-year-old male	15 adenomas	Yes	
Halo36	67-year-old male	At least 10 adenomas	Yes	
Halo40	70-year-old female	18 adenomas, 3 HPPs Carrier of monoallelic *MUTYH* mutation (*MUTYH* c.920G>A, pR307G)	Yes	
Halo41	51-year-old male	10 adenomas, 10 HPPs Carrier of a monoallelic *MUTYH* mutation (*MUTYH* c.1187G>A G396D)	Yes	
Halo43	67-year-old male	Referred from bowel screening. Colonoscopy showed 18 polyps and CRC. It is not clear how many polyps were examined microscopically (at least 5) but these all showed adenoma. The CRC was moderately differentiated ?stage	No	
Halo44	72-year-old female	8 TAs, 2 HPPs, one 'mixed HPP/ TA'	Yes	
Halo45	77-year-old male	27 adenomas, 7 HPP	Yes	
Halo46	44-year-old female	Clinical FAP – has had a colectomy and proctectomy	Yes	
Halo47	51-year-old female	Colonoscopy found in excess of 50 sessile polyps throughout the colon. 2 of these were biopsied: TVA. The patient had a subtotal	Yes	Halo68 (sister): 54-year-old female. 14 TA

		colectomy - ?4 polyps were sampled: 2 TA, 2 serrated adenomas. Subsequent completion colectomy identified a CRC (pT4N2Mx)		with LGD and 6 HPPs
Halo48	71-year-old male	Referred from screening. 30 polyps: 4 have been biopsied: all adenoma	Yes	
Halo49	57-year-old female	Rectal cancer age 32 (?stage), ameloblastoma age 38. 22 polyps: mixture of adenomas and hyperplastic lesions	No	
Halo50	64-year-old male	10 adenomas	No	
Halo51	43-year-old female	208 adenomas	Yes	Halo77 (mother), Halo78 (father): unaffected Both are described as unaffected by St. Mark's Hospital, where a colonoscopy was presumably performed, but this is unconfirmed. RNA available for both
Halo52	59-year-old male	Over 1000 adenomas	Yes	

Halo53	51-year-old male	Approximately 400 adenomas, CRC age 27 (Duke's C)	Yes	Halo75 (mother), Halo76 (father): unaffected. Both have had negative colonoscopies. RNA available for both
Halo54	70-year-old male	Approximately 260 polyps, rectal cancer age 46 (Duke's A)	No	
Halo55	54-year-old male	18 polyps: 9 adenomas, 9 serrated lesions	Yes	M_Halo55 (mother), F_Halo55 (father): unaffected according to clinical history. RNA available for both
Halo56	38-year-old female	57 adenomas	Yes	Halo73 (mother), Halo74 (father): unaffected. Mother had one polyp in 2002 but 2 clear colonoscopies subsequently. Father had 2 polyps at colonoscopy

Halo57	40-year-old male	Approximately 3500 adenomas	No	
Halo58	64-year-old male	Approximately 100 polyps: mixture of adenomas and HPPs	Yes	
Halo61	60-year-old female	11 polyps: 2 SAs, 7 TA, 1 TVA, and 1 HPP	Yes	
Halo62	76-year-old female	Ileorectal anastomosis, but the indication for prior surgery is unclear. 10 rectal adenomas	Yes	
Halo63	68-year-old male	At least 24 polyps. 5 biopsies show 4 adenomas and 1 HPP	Yes	
Halo64	54-year-old female	Thousands of colorectal polyps. Those which have been biopsied showed adenomas. CRC aged 23 (?stage)	Yes	
Halo65	33-year-old male	117 adenomas	Yes	
Halo66	70-year-old male	23 adenomas, 4 HPPs	Yes	
Halo67	73-year-old male	At least 15 adenomas	Yes	
Halo69	73-year-old female	At age 72 had CRC, pT4, and at least 24 polyps. It is not clear how many were removed, but they included adenomas, HPPs and one SA	No	
Halo70	71-year-old male	15 polyps. 11 shown to be adenomas on microscopy	Yes	

Halo71	65-year-old female	14 adenomas	Yes	
Halo72	61-year-old male	At least 10 adenomas, one of which contained moderately differentiated CRC (?stage). Additionally, had Duke's A sigmoid carcinoma age 53	Yes	
Halo79	80-year-old male	Colectomy for polyposis. No other information	No	
Halo80	76-year-old female	38 adenomas	Yes	
Halo81	65-year-old male	17 polyps: 8 adenomas (4 TA with LGD, 3 TVA with LGD, 1 TVA arising in a SSL) and 9 HPPs	No	

Table 2.1: Table describing the demographic details and clinical phenotype of the NMI polyposis patients and their relatives included in this study

Unique Identifier	Sample Details	Genotype	RNA	Further Information
Halo01	Control FAP	*APC* c.3631-3632delAT	No	
F014_M_004_1	Control FAP	*APC* c.2805 C>A	No	
F014_M_005_1	Control FAP	*APC* c.4393-4394delGA	No	
F014_M_006_1	Control FAP	FAP mosaic: 5% mosaic *APC* c.4393-4394delGA	No	
Halo10	Control FAP	*APC* c.3183_3187delACAAA	No	
Halo12	Control FAP	*APC* c.3927-3931delAAAGA	No	
Halo16	Control FAP	*APC* c.3408delA	No	
FAPPol51	Control FAP	*APC* promoter 1B deletion	Yes	
FAPPol141	Control FAP	Deletion of *APC* exons 11 and 12	Yes	
Halo42	Control FAP, monoallelic *MUTYH* mutation	*APC* c.1187dupA, *MUTYH* c.536A>G	Yes	
Halo37	Control MAP	Compound heterozygous for *MUTYH*	Yes	

		c.536A>G and c.649C>T		
Halo39	Control MAP	Compound heterozygous for *MUTYH* *c.*303G>T and c.312C>A	No	
Halo59	Control MAP	*MUTYH* c.1187G>A homozygote	Yes	
MAPPol71	Control MAP	Compound heterozygous for *MUTYH* c.303G>T and c.312C>A	Yes	
MAPPol90	Control MAP	Compound heterozygous for *MUTYH* c.1187G>A and c.536A>G	Yes	
Halo03	89-year-old male Healthy control		No	Father of a patient with jaw osteomas recruited to the study but not included in this thesis
Halo11	79-year-old female Healthy control		No	Mother of a patient with jaw osteomas recruited to the study but not included in this thesis

Halo95	62-year-old female Healthy control		Yes	Mother of a patient with HPPs recruited to the study but not included in this thesis

Table 2.2: Table describing the sample details and the genotype of control samples included in this study.

2.3 General Techniques

2.3.1 Agarose Gel Electrophoresis

Agarose gels were made with 1x TAE buffer to a concentration of 0.8-1.5%. The appropriate mass of agarose was added to 100 ml of 1x TAE and heated in a microwave for 2 minutes to allow it to melt. The liquid was cooled by running it under a cold tap for 2-5 minutes. 5 µl of Gel Red was added, and the liquid poured into a gel mould to set. The gel was completely submerged in 1x TAE buffer in an AB0708 100 V gel tank or a Cleaver Scientific gel tank. Unless otherwise stated, 5 µl of Orange G loading dye was added to 5 µl of sample, and the entire volume was loaded onto the gel. A 1Kb plus DNA ladder (ThermoFisher) was loaded alongside the samples. Electrophoresis was performed at 100 V for 30 minutes using the BioRad 200/2.0 power supply. Following electrophoretic separation of the samples, visualisation was achieved using the BioDoc-It Imaging System Benchtop UV Transilluminator.

2.3.2 Germline DNA Extraction

Genomic DNA was extracted from venous blood using the Chemagic Star Automated Extraction Method by the All Wales Medical Genetics Service (AWMGS). At least 4 ml of peripheral blood was collected in an EDTA tube. The blood was then transferred into a 50 ml tube and lysis buffer added for up-front DNA extraction. The Chemagic Star instrument uses polyvinyl alcohol particles (M-PVA Magnetic Beads), which have a hydrophilic surface, to bind to nucleic acids. Once bound, the nucleic acids were transferred with a magnetisable rod to different wash buffers and finally to an elution buffer

(http://www.hamiltonrobotics.com/fileadmin/user_upload/Standard_Solutions/B-1110-03_Chemagic_STAR_web.pdf. Accessed 21/5/14). DNA was stored at -20^{0}C.

2.3.3 RNA Extraction

RNA was extracted from whole blood using a Tempus Spin RNA Isolation Reagent Kit (Applied Biosystems). 3 ml peripheral blood was collected in a Tempus blood RNA tube (Life Technologies) and shaken vigorously for 30 seconds. The blood sample was poured into a clean 50 ml conical tube. 1x Phosphate Buffered Saline (PBS) was added to bring the volume up to 12 ml. The tube was vortexed for 30 seconds, then centrifuged at 4^{0}C for 30 minutes at 3000 x g. The supernatant was poured into a disinfectant solution. The tube was inverted on absorbent paper for 1 to 2 minutes, and any remaining liquid was blotted from the rim of the tube with absorbent paper. 400 µl of RNA purification re-suspension solution was added to the tube, and the tube was vortexed briefly to re-suspend the RNA pellet. RNA was then purified according to the manufacturer's protocol. Briefly, this involved pre-wetting a filtration membrane with RNA purification wash solution 1, pipetting the re-suspended RNA onto the filter, centrifuging the sample for 30 seconds at 16 000 x g, discarding the liquid waste, pipetting 500 µl wash solution 1 onto the filter and centrifuging. The liquid waste was discarded, then the process was repeated twice with wash solution 2. The filter was centrifuged to dry it, then 90 µl elution solution was added. The tube underwent centrifugation, and the eluted RNA was subsequently pipetted back onto the filter, which was centrifuged again. All centrifugation steps were performed at 16 000 x g for 30 seconds, apart from the final centrifugation, which was carried out for 2 minutes. Eluted RNA was transferred to a fresh tube and stored at -80^{0}C.

2.3.3.1 Assessment of RNA Quality

RNA integrity was assessed by electrophoresing 3 µl RNA, along with 8 µl Orange G loading dye, though a 0.8% agarose gel for 40 minutes at 120 V. Images of the gel were taken using the BioDoc-It Imaging System Benchtop UV Transilluminator. Any degraded samples were omitted from further analysis. If the RNA sample

contained a significant amount of genomic DNA, it was omitted from further analysis.

2.3.3.2 RNA Conversion to cDNA

Following RNA extraction, conversion to cDNA was carried out by making a 12 µl reaction mix containing 2 µl random primers (100 ng/µl), 1 µl dNTPs (10 mM) and 9 µl RNA (2.5 µg/µl). This was incubated for 5 minutes at 65^{0}C on the G Storm to disrupt any template secondary structures. It was then placed on ice to prevent the re-formation of secondary structures. To the reaction mixture, 1 µl of reverse transcriptase was added (Superscript II 200 U/µl), and the mixture incubated for 2 minutes at 25^{0}C on the G Storm. Subsequently, 4 µl 5x first strand buffer, 2 µl 0.1 M DTT and 1 µl RNAsin was added, and the reaction mix was incubated for 10 minutes at 25^{0}C, 50 minutes at 42^{0}C and 15 minutes at 70^{0}C on the G Storm. cDNA was stored at -20^{0}C.

2.3.4 DNA Extraction from FFPE Tissue

To extract DNA from FFPE tissue, the GeneRead FFPE Kit (Qiagen) was used. 10 µm tissue sections were cut by the Cellular Pathology Department, University Hospital of Wales. Excess paraffin was trimmed using a sterile scalpel, and the tissue was placed in a 1.5 ml Eppendorf tube. 160 µl deparaffinisation solution was added, and the tube was vortexed vigorously for 10 seconds then centrifuged briefly to bring the sample to the bottom of the tube. DNA extraction was carried out following the manufacturer's protocol. Briefly, this involved incubating the solution at 56^{0}C for 3 minutes, then allowing it to cool to room temperature. 55 µl RNase-free water, 25 µl Buffer FTB, and 20 µl proteinase K were added to the solution, which was vortexed and briefly centrifuged. The solution was incubated at 56^{0}C for 1 hour, then at 90^{0}C for 1 hour. The solution was briefly centrifuged to remove drops from inside the lid. The lower, clear phase was transferred to a new microcentrifuge tube. 115 µl RNase-free water was added and the solution mixed. 35 µl UNG was added, and the solution was vortexed and incubated at 50^{0}C for 1 hour in a thermomixer. Again, the mix was briefly centrifuged to remove drops from inside the lid. 2 µl RNase A (100 mg/ml) was added, and the solution was mixed and incubated for 2 minutes at room temperature. 250 µl Buffer AL was added to the sample, which was mixed thoroughly by vortexing. 250 µl ethanol (96–100%)

was added and the mixture vortexed and centrifuged. Subsequently 700 µl of lysate was transferred to the QIAamp MinElute column, in a 2 ml collection tube. The lid was closed and the tube centrifuged. The flow-through was discarded and the collection tube re-used. The process of transferring the lysate/ centrifugation/ discarding the flow-through was repeated until the lysate was used up. 500 µl Buffer AW1 was added to each spin column and the mixture centrifuged. The flow-through was discarded and the collection tube re-used. The process was repeated using 500 µl Buffer AW2, then 250 µl ethanol. The spin column was then placed into a new 2 ml collection tube and centrifuged to remove any residual liquid. The QIAamp MinElute column was transferred into a clean 1.5 ml microcentrifuge tube. 40 µl Buffer ATE was pipetted on to the centre of the membrane. The tube was incubated at room temperature for 5 minutes, then centrifuged. All centrifugations were performed for 1 minute at full speed. DNA was stored at -20^{0}C.

2.3.5 Quantification of Nucleic Acids

Two platforms were used to quantify nucleic acids – the Nanodrop and the Qubit 2.0 Fluorometer. In both cases, 2 µl of sample was used for quantification, following the manufacturer's protocol.

2.3.6 Primer Design

Unless otherwise stated, all primers were designed using Primer3 v0.4.0 (Rozen and Skaletsky 2000) and were supplied by Eurogentec or Eurofins. The specificity of primer pairs was determined using UCSC In Silico PCR software (https://genome.ucsc.edu/cgi-bin/hgPcr).

2.3.7 Bisulfite Conversion of DNA

Sodium bisulfite conversion of DNA is a technique which allows the detection of unmethylated versus methylated cytosines. It involves the deamination of unmodified cytosines to uracil, leaving the methylated cytosines intact.
DNA was bisulfite-converted using the Epitect Bisulfite Kit (Qiagen) following the manufacturer's protocol. Briefly, this involved making a mixture of 2 µg DNA, 85 µl Bisulfite Mix, 35 µl DNA protect buffer and an appropriate volume of RNase-free

water to make the total volume 140 µl. The mixture was vortexed. The bisulfite conversion was carried using the G Storm thermal cycler for 5 minutes at 95^{0}C, 25 minutes at 60^{0}C, 5 minutes at 95^{0}C, 85 minutes at 60^{0}C, 5 minutes at 95^{0}C and 175 minutes at 60^{0}C. The PCR tubes were briefly centrifuged, then the complete reaction was transferred to a clean tube. 560 µl Buffer BL containing 10 µg/ml carrier RNA was added to each sample. This was then vortexed and briefly centrifuged. The entire reaction mix was transferred into an Epitect spin column, which was centrifuged at maximum speed for 1 minute. The flow-through was discarded and the spin column placed back in the collection tube. The process was repeated with 500 µl Buffer BW. 500 µl Buffer BD was added and the mixture incubated for 15 minutes at room temperature. The spin columns were centrifuged at maximum speed for 1 minute. The flow-through was discarded, and the spin columns were placed back in the collection tubes. The process was repeated two further times, using 500 µl Buffer BW. The spin columns were then placed in clean collection tubes and centrifuged at maximum speed for 1 minute to remove any residual liquid. The spin columns were placed with the lids open into clean microcentrifuge tubes and incubated for 5 minutes at 56^{0}C in a heating block. Spin columns were then placed into clean microcentrifuge tubes. 20 µl Buffer EB was pipetted onto the centre of each membrane, and the purified DNA was eluted by centrifugation for 1 minute at 15 000 x g. To complete the process, 20 µl of Buffer EB was pipetted to the centre of each membrane and centrifuged for 1 minute at maximum speed. The resultant bisulfite-converted DNA was stored at -20^{0}C.

2.3.8 Standard PCR and Sanger Sequencing

The standard steps for PCR and Sanger Sequencing are:

1. PCR
2. Gel Electrophoresis
3. ExoSap PCR Purification
4. Big Dye Reaction
5. Isopropanol Clean Up
6. Sequencing and Data Analysis

These shall be described in sections 2.3.8.1-2.3.13.

2.3.8.1 Standard Reagents for PCR

Water	11 µl
Forward primer (10 µM)	0.25 µl
Reverse primer (10 µM)	0.25 µl
Megamix Gold (MMG)	12.5 µl
DNA/ cDNA (5 ng/µl)	1 µl

2.3.8.2 Reagents for PCR for FFPE Tissue DNA

Water	Variable
Forward primer (10 µM)	0.25 µl
Reverse primer (10 µM)	0.25 µl
Megamix Gold (MMG)	15 µl
DNA	10 ng

2.3.8.3 Reagents for *Fast*-COLD-PCR

Water	11 µl
Forward primer (10 µM)	0.25 µl
Reverse primer (10 µM)	0.25 µl
Megamix Gold (MMG)	12.5 µl
DNA (5 ng/µl)	1 µl

2.3.8.4 Reagents for *Full*-COLD-PCR

Water	11 µl
Forward primer (10 µM)	0.25 µl
Reverse primer (10 µM)	0.25 µl
Megamix Gold (MMG)	12.5 µl
DNA (5 ng/µl)	1 µl

2.3.8.5 Standard PCR Reaction Conditions:

95^{0}C for 5 minutes

95^{0}C for 1 minute

Annealing temperature: 58^{0}C for 1 minute

Elongation temperature 72^{0}C for 1 minute

Number of cycles: 35

72^{0}C for 5 minutes

2.3.8.6 Reaction Conditions for *Fast*-COLD PCR:

95^{0}C for 5 minutes

70^{0}C - 95^{0}C for 1 minute

58^{0}C for 1 minute

72^{0}C for 1 minute

Number of cycles: 35

72^{0}C for 5 minutes

2.3.8.7 Reaction Conditions for *Full*-COLD-PCR:

Cycle 1 95^{0}C for 5 minutes

95^{0}C for 1 minute

58^{0}C for 1 minute

72^{0}C for 1 minute

Number of cycles: 10

72^{0}C for 5 minutes

Cycle 2: 95^{0}C for 1 minute

70^{0}C for 5 minutes

70-95^{0}C for 1 minute

58^{0}C for 1 minute

72^{0}C for 1 minute

Number of cycles: 25

72^{0}C 5 for minutes

For both the *fast* and the *full* protocols, the initial COLD-PCR protocol allowed the determination of the lowest denaturation temperature over a wide range of temperatures (70-95°). Once this was ascertained, the protocol was repeated using a narrower temperature gradient to further specify the lowest denaturation temperature.

2.3.9 Gel Electrophoresis

To confirm that PCR reactions had worked and that the PCR products were the correct size, products were electrophoresed through a 1.5% agarose gel, with a current of 120 V for 30 minutes. Images of the gel were taken using the BioDoc-It Imaging System benchtop UV transilluminator (see 2.3.1).

2.3.10 ExoSap PCR Purification

The ExoSap PCR purification reaction involves an exonuclease enzyme degrading excess primers and any ssDNA present, and an alkaline phosphatase enzyme degrading any dNTPs. The ExoSap stock was made from Exonuclease and Shrimp Alkaline Phosphatase at a ratio of 1:2. 1 µl of ExoSap was added to PCR products. These were incubated on the G Storm for one hour at 37^0C , then for 15 minutes at 85^0C .

2.3.11 Big Dye Reaction

The Big Dye Terminator v3.1 Sequencing Kit was used for the sequencing reaction. 1 µl of PCR product was added to a reaction mixture containing:

Big Dye v.3.1	0.75 µl
Water	5.25 µl
Primer (10 µM)	1 µl
BD Buffer	2 µl

This then underwent thermal cycling on the G Storm: 25 cycles of 95^0C for 10 seconds, 50^0C for 5 seconds, 60^0C for 3 minutes and 30 seconds.

2.3.12 Isopropanol Clean-Up Protocol

Isopropanol was used to precipitate DNA to allow purification. 40 µl of 75% isopropanol was added to each sample, which were incubated for 30 minutes at room temperature before being centrifuged for 45 minutes at 4000 revolutions per minute (rpm). The samples were inverted, placed into the centrifuge upside down and spun for 30 seconds at 500 rpm. Once this was complete, they were left in the dark to dry for 10 minutes before 10 µl of Hi-Di was added to re-suspend the DNA.

2.3.13 Sequencing and Data Interpretation

The samples were sequenced on the ABI Prism 3100 Genetic Analyser. Sequencing data was read using Sequencher 5.2.4 software.

2.3.14 Methylation-Specific PCR (MSP)

Methylation-specific PCR is a technique used to detect DNA methylation. Primers are designed to bind to bisulfite-converted DNA dependent upon whether the original DNA molecule was methylated/ unmethylated. The presence or absence of a PCR product will therefore indicate the methylation status of the original DNA. In this thesis, MSP was used to determine the methylation status of the *APC* 1A and 1B promoters.

2.3.14.1 Reagents for *APC* Promoter 1A MSP

Epitect Mastermix	25 µl
Forward primer	500 ng
Reverse primer	500 ng
Water	Variable
DNA	50 ng
Total:	50 µl

2.3.14.2 Reaction Conditions for *APC* Promoter 1A MSP

95^0C	10 minutes

35 cycles of:

95^{0}C	30 seconds
53.9^{0}C	30 seconds
72^{0}C	30 seconds

72^{0}C 10 minutes

2.3.14.3 Reagents for *APC* Promoter 1B MSP

Epitect Mastermix	25 µl
Forward primer	250 ng
Reverse primer	250 ng
Water	Variable
DNA	100 ng
Total:	50 µl

2.3.14.4 Reaction Conditions for *APC* Promoter 1B MSP

95^{0}C	10 minutes

35 cycles of:

95^{0}C	30 seconds
50.1^{0}C	30 seconds
72^{0}C	30 seconds

72^{0}C 10 minutes

2.3.15 Quantitative PCR (qPCR)

Quantitative PCR (qPCR) is a PCR technique in which the reaction is monitored in real time. This allows the quantification of the input sample, as higher sample concentrations will result in PCR products which can be detected at an earlier stage. This study used Taqman technology (ThermoFisher). The principles underpinning it are that sample DNA is denatured, at the appropriate annealing temperature primers bind, along with a Taqman probe which includes a fluorescent dye and a non-fluorescent quencher (NFQ). During DNA synthesis, using the primers and the

template DNA, the 5'nuclease activity of the *Taq* polymerase will cleave the probe, separating the dye from its quencher. The dye will then emit a signal which is detected. With each cycle of PCR amplification, increasing amounts of dye are released, therefore the fluorescent signal will increase.

Prior to performing qPCR reactions, the basic foundations for this work were established, i.e.

1. Determination of optimal RNA input
2. Determination of reaction efficiencies, to ensure that that ΔΔCt method could be used for data analysis

The details of this are described in Appendix 2.1.

2.3.15.1 qPCR Reagents

20X Taqman Gene Expression Assay	1 µl
2X Taqman Gene Expression Master Mix	10 µl
cDNA template	4 µl
Water	5 µl

2.3.15.2 qPCR Reaction Conditions

Hold 50^0C for 2 minutes
Hold 95^0C for 10 minutes
Cycle (40 cycles): 95^0C for 15 seconds, 60^0C for 1 minute

The thermal cycler used was the QuantStudio 12K Flex Real-Time PCR System (Life Technologies). qPCR reactions were carried out in triplicate for each sample. A no-template control was included for each assay.

2.3.16 *APC* and *MUTYH* Expression in Healthy Controls

This study involved comparing the cDNA levels of *APC* and *MUTYH* in NMI polyposis patients to a cohort of healthy controls.

2.3.16.1 qPCR in a Healthy Control Cohort

Venepuncture was performed on the healthy control cohort after each volunteer had given informed consent. RNA was extracted (2.3.3, 2.3.3.1) and reverse transcribed to cDNA (2.3.3.2). A total of 44 cDNA samples from healthy relatives and healthy controls underwent qPCR. The reagents and reaction conditions are described above (2.3.15.1, 2.3.15.2). The Taqman assays used are listed in Table 2.3:

Gene	**Designation**	**Taqman Assay**
GAPDH	Endogenous control	HS02758991_g1
ACTB	Endogenous control	HS99999903_m1
APC	Target	HS01568269_m1
MUTYH	Target	HS01014856_m1

Table 2.3: Taqman assays used for qPCR

2.3.16.2 Data Analysis

Results were interpreted using ThermoFisher Cloud software (https://apps.thermofisher.com/apps/dashboard/#/). The first stage of analysis was a quality assessment. Any results which were flagged as outliers were examined. This usually referred to results which had a CT value which deviated significantly from the CT values in the associated replicate group, or whole replicate groups which had a significant deviation within them.

The individual replicate results which deviated significantly from their replicate group were excluded as this was likely due to experimental error and would unfairly skew results. Those replicate groups which had a significant deviation within them were excluded as the results were not tight enough for robust analysis. If a sample had a complete replicate group for one of the endogenous controls excluded, that sample was removed from further analysis as we were seeking to normalise results to two different endogenous controls. Therefore, some samples needed to undergo qPCR a second time to obtain usable results.

The final 40 samples included in this study are listed in the Table 2.4:

Sample ID	Sample group
C2-C27, C30-C33, C35, C37, C39	Healthy Controls
M_Halo55	Healthy Controls
F_Halo55	Healthy Controls
Halo75	Healthy Controls
Halo76	Healthy Controls
Halo77	Healthy Controls
Halo78	Healthy Controls
Halo79	Healthy Controls

Table 2.4: Healthy control samples used for qPCR studies

2.4 Next Generation Sequencing (NGS)

2.4.1 Haloplex Assay

The probes for the Haloplex assay were designed using Agilent's Sure Design software (https://earray.chem.agilent.com/suredesign/home.htm). The probes generate read lengths of 250 base pairs.

2.4.2 Target Gene Capture and Sequencing

The concentration of the DNA samples to be examined were determined using the Qubit 2.0 Fluorometer. The DNA was diluted to a concentration of 5 ng/µl. Sequencing was carried out by Dr. James Colley, WGP, using the HiSeq 2500 (Illumina). The protocol involved digesting gDNA with 16 different restriction enzymes to create a library of restriction fragments. Digested DNA was hybridised to Haloplex probes for target enrichment and sample indexing. The DNA-Haloplex probe hybrids were captured by streptavidin beads, prior to the ligation of the circularised fragments. DNA was eluted with sodium hydroxide and underwent PCR to amplify the target libraries. Target libraries were purified, quantified and pooled before sequencing.

2.4.3 Bioinformatic Analysis Following UDS

Bioinformatic analysis was performed by Dr. Peter Giles, WGP.
Reads were analysed using the following procedure:

1. Raw reads (in fastq format) were transferred from the sequencer to the local computer cluster
2. Reads were mapped against the hg19 genome using bwa-mem, producing a BAM file
3. Quality control was reviewed using 10% of reads using a combination of in-house scripts, FastQC and circos for visualisation
4. GATK was then used to recalibrate quality scores and to apply a localised realignment around indels before filtering for artefactual duplicates using samtools
5) Variant calling was performed using GATK (Unified Genotyper) and varscan (for low frequency variants) producing VCF files
6) Variants were annotated, including the following parameters:
- Coverage
- Allele balance
- UCSC genes
- 1000 genomes
- dbsnp137
- Gene description
7) Variants were visualised using Integrative Genomics Viewer (IGV)

2.4.4 Whole Exome Sequencing (WES)

Sequencing was performed by the WGP. Briefly, 50 ng of gDNA was used as the input template. Sequencing libraries were prepared using the Illumina Nextera Rapid Capture Enrichment kit. Subsequent steps included tagmentation of the gDNA, clean-up of the tagmented DNA, amplification of DNA, clean-up of amplified DNA, hybridisation of probes, capture of the hybridised probes, second hybridisation of probes, second capture, clean-up of the captured library, amplification of enriched library, clean-up of the enriched library, validation of the complete library.
The manufacturer's instructions were largely followed with extra quantitation steps prior to the hybridisation of the probes to ensure that close to 50 ng of each sample was pooled. The libraries were validated using the Agilent 2100 Bioanalyser and a high-sensitivity kit (Agilent Technologies) to ascertain the insert size, and the Qubit

2.0 Fluorometer was used for quantitation. Following validation, the libraries were normalised to 4 nM, pooled together and clustered on the cBot™2 following the manufacturer's recommendations. The pool was sequenced using a 75-base paired-end (2x75bp PE) dual index read format on the Illumina HiSeq 2500 in high-output mode according to the manufacturer's instructions.

2.4.5 Bioinformatic Analysis Following Whole Exome Sequencing

Bioinformatic analysis was performed by Dr. Kevin Ashelford (WGP).
Reads were analysed following the steps outlined below:

1. The raw output from the HiSeq 2500 were BCL files. These, once transferred from the sequencing rig, were converted to fastq.gz using the Illumina tool bcl2fastq (version 2.17.1.14; https://support.illumina.com/content/dam/illumina-support/documents/documentation/software_documentation/bcl2fastq/bcl2fastq_letterbooklet_15038058brpmi.pdf)
2. Initial quality control was performed on 10% of the data (for speed) using a combination of an in-house script and FastQC (version 0.11.2; http://www.bioinformatics.babraham.ac.uk/projects/fastqc/). The in-house script performed a quick mapping of the data using BWA-MEM (version 0.7.4; http://bio-bwa.sourceforge.net) and then calculated percentage mapped and insert size
3. Reads were trimmed with Trimmomatic (version 0.35; http://www.usadellab.org/cms/?page=trimmomatic) to remove any adapter and poor-quality base calls from the ends of reads
4. Trimmed reads were then mapped with BWA-MEM (version 0.7.10). Mapping was performed against standard human reference hg19 (Feb. 2009 (GRCh37/hg19) — http://genome-euro.ucsc.edu/cgi-bin/hgGateway?db=hg19&redirect=manual&source=genome.ucsc.edu)
5. Mapping outputs were then post-processed using several GATK (version 3.3.0; https://software.broadinstitute.org/gatk/) tools: (i) RealignerTargetCreator and IndelRealigner to realign reads around indels, (ii) BaseRecalibrator and PrintReads to recalibrate raw Quality Values. The Picard toolkit (version 1.118; https://broadinstitute.github.io/picard/) tool MarkDuplicates was then used to flag duplicate reads

6. After post-processing a further round of quality control was performed using FastQC as before and an in-house script to calculate on-target mapping metrics
7. Raw variant calls were called on each sample separately using GATK HaplotypeCaller in GVCF mode followed by GenotypeGVCFs to produce a final combined vcf file. Raw variant calls were then filtered using VariantRecalibrator to provide Variant Quality Score Recalibration (VQSR) which allows for dynamic filtering of variants using machine learning techniques
8. Variants were then annotated against databases dbSNP version 138, 1000Genomes, COSMIC and SIFT using ANNOVAR (version 20150322; http://annovar.openbioinformatics.org). In-house scripts were used to format the output and extract additional filtering criteria including coverage depths, allele balance, and gene descriptions

2.4.6 Single Molecule Molecular Inversion Probe Gene Capture, Sequencing and Bioinformatic Analysis

1. A reaction mixture composed of the following reagents was prepared:
    - DNA (5 ng/µl) 10.0 µl
    - Hemo Klen Taq DNA Polymerase (10 U/µl) 0.32 µl
    - Ampligase DNA ligase (100 U/µl) 0.01 µl
    - smMIP 3.49 µl
    - 10X Ampligase DNA ligase buffer 2.5 µl
    - dNTPs (0.25 mM) 0.32 µl
    - Water 8.36 µl

    To make the reaction mix, the ligase, buffer and dNTPs were combined. To this, the smMIPs and water were added, then finally the DNA
2. The mix was incubated at 95^{0}C for 10 minutes, and then at 60^{0}C for 18 hours
3. To ensure that the gene capture was successful, 5 µl of the reaction mixture was electrophoresed through a 3% agarose gel for 30 minutes. Images of the gel were taken with the Bio Rad Gel Doc XR+ System
4. The reaction mixture was cooled and kept on ice. An exonuclease reaction mixture was added to each sample. This was composed of:
    - EXOI 0.5 µl
    - EXOIII 0.5 µl
    - Ampligase buffer 0.2 µl

- Water 0.8 µl

5. The mixture was incubated at 37^{0}C for 45 minutes and then at 95^{0}C for 10 minutes
6. A test PCR was carried out on a subset of the reactions, to ensure that PCR amplification would be successful. The following reagents were used:
    - 2X iProof HF Mastermix 12.5 µl
    - Illumina forward primer* (100 µM) 0.125 µl
    - Barcoded reverse primer PGM* (100 µM) 0.125 µl
    - Water 7.25 µl
    - Template MIP reaction 5 µl

(*Test PCR primers: SLXA_PE_MIPBC_FOR: AATGATACGGCGACCACCGAGATCTACACATACGAGATCCGTAATCGGGAAGCTGAAG; SLXA_PE_MIPBC2_REV_001: CAAGCAGAAGACGGCATACGAGAT**GTTAAGAC**ACACGCACGATCCGACGGTAGTGT **Bold:** sample barcode)

7. The reaction mixture underwent the following reaction:
    - 98^{0}C for 30 seconds

    Cycles of: tumour tissue: 24 cycles, normal tissue 21 cycles:
    - 98^{0}C for 10 seconds
    - 60^{0}C for 30 seconds
    - 72^{0}C for 2 seconds
8. 5 µl of PCR products were electrophoresed through a 3% agarose gel for 30 minutes. The PCR bands were visualised using the Bio Rad Gel Doc XR+ System. Provided that PCR had been successful, all samples underwent the reaction. The reagents for each reaction were as follows:
    - 2X iProof HF Mastermix 37.5 µl
    - Illumina forward primer* (100 µM) 0.375 µl
    - Barcoded reverse primer PGM* (100 µM) 5.0 µl
    - Water 18.375 µl
    - Template MIP reaction 15 µl
9. Reaction conditions are as described in Step 7
10. 5 µl of PCR products were electrophoresed through a 3% agarose gel for 30 minutes. The PCR bands were visualised using the Bio Rad Gel Doc XR+ System
11. Each sample was given a score based on the intensity of its PCR band: High, Medium, Low

12. The samples were pooled so that approximately equivalent amounts of each sample were included. There were separate pools for tumour DNA and germline DNA
13. The pools were run on a 3.5% agarose gel for 30 minutes. Images of the gel were taken with the Bio Rad Gel Doc XR+ System
14. The PCR products were extracted from the gel using a Zymoclean Gel DNA Recovery Kit
15. PCR products were assessed using the TapeStation to confirm that products of the correct size were present. If the TapeStation was suggestive of >1 product in the reaction mix, steps 13-15 were repeated to try and purify the desired product
16. The samples were quantified using the Qubit Fluorometer 2.0, following the manufacturer's protocol
17. All pools were combined into a megapool, and the concentration was quantified using the Qubit Fluorometer 2.0
18. The megapool was diluted to 2nM
19. Sequencing was performed using the MiSeq (Illumina)

Following sequencing, bioinformatic analysis was carried out by Laura Chegwidden, Oxford University/ Birmingham University. The protocol is outlined below:

Samples were demultiplexed using basespace. The FASTQ files were downloaded for each sample. Each FASTQ file was processed using Perl scripts from Roland Arnold to:

1. remove MIP backbones from read 1 (R1) and read 2 (R2)

2. remove reads that were <55bp long after processing

3. remove reads whose mate had been filtered out to leave paired reads only

Next, tools were used as suggested by the MIPGEN pipeline builder (https://github.com/shendurelab/MIPGEN/tree/master/tools):

1. Use PEAR to join R1 and R2
2. Use a python script to move the single molecule tag to the FASTQ header

Mapping and basic quality control was performed following the processes described below:

1. BWA mem was used to align reads to hg19
2. Python script was employed to collapse reads into single molecule tag defined read groups
3. Statistics were generated on the number of panel targets covered at 20X to identify samples that had failed. A sample was considered as having "passed" basic sequence QC if >50% of targets were covered at 20x

Samtools was used to make BAM files.

For Indel realignment, pre-processing was carried out as per the lofreq website. This involved Viterbi realignment, adding indel quality scores to files and adding alnqual. Variants were called using Lofreq-star:

- call-indels
- m 60 (filtering on mapping quality >60)
- no-default-filter (these were set at a later stage)

Following calling, variants were filtered according to the following criteria:

- cov-min 10 (min coverage 10)
- sb-thresh 60 (strand bias phred 60)
- a 0.01 (MAF>0.01)

Formatting/ further filtering involved:

- Filtering the data so that calls with >5 reads in either Forward OR Reverse remained. Those with <5 reads were omitted
- Assigning genotypes (VAF<0.75=HET; VAF>=0.75=HOM)
- MAF>0.05
- Assessing strand bias: if there were reference reads in Forward/Reverse, but there were no variant reads in the opposite strand, the variant was omitted
- Annotation using VEP
- Removing likely germline variants
- Filter "ExAC.r0.3.nonTCGA_AF > 0.01"

- Filter "UK10K_AF > 0.01"
- Filter "EUR_MAF > 0.01"
- Adjusting formats & layouts so that vcf (original lofreq call) columns were reported first, then VEP annotations
- Filtering to remove intronic, synonymous and splice region variants
- Variants were removed if they were part of a homopolymer repeat >4 in length
- Variants were removed if they were discovered as part of a search for panel artefacts- poorly performing molecular inversion probe/ hard to sequence region
- The high confidence call sets had a depth > 20 and VAF support > 5%

2.5 Molecular Biology Techniques

2.5.1 Plasmid Retrieval from Filter paper

The demarcated spot containing plasmid was cut out from the filter paper using a clean razor blade. It was immersed in 30 µl TE and mixed with a pipette. It was left for at least 10 minutes at room temperature prior to bacterial transformation. If it was not used immediately, DNA was stored at -20°C.

2.5.2 Preparation of Lysogeny Broth (LB)

To prepare LB, 20 g LB Broth Base was added to 1 L distilled water. It was autoclaved at 126^0C for 45 minutes. An appropriate antibiotic, at 1:1000 dilution, was added when the solution had cooled down. The broth was stored at room temperature.

2.5.3 Preparation of Lysogeny Broth (LB) Agar

To prepare LB agar, 10.5 g LB Agar was added to 300 ml distilled water. It was autoclaved at 126^0C for 45 minutes. It was allowed to cool slightly, prior to adding an appropriate antibiotic, at 1:1000 dilution, when hand-warm. It was poured into

plates next to a naked flame. Once the agar had set, it was turned upside down to prevent condensation from dripping onto the agar. The plates were stored at 4^0C .

2.5.4 Bacterial Transformation of XL1-Blue Competent Cells

All steps of the protocol were carried out next to a naked flame, and the equipment openings/ lids were quickly waved through the flame. Bacterial transformation followed the protocol detailed below:

1. The waterbath was warmed to 42^0C
2. SOC medium was warmed to room temperature
3. The agarose plates containing the desired antibiotic were warmed to 37^0C in an incubator
4. Competent *E. Coli* cells were thawed on ice
5. 25 µl bacteria was transferred into a pre-cooled 1.5 ml Eppendorf tube
6. 0.4 µl β-mercaptoethanol was added to each tube, which were kept on ice
7. 1 µl circularised DNA was added and the solution and mixed gently
8. The mix was incubated on ice for 30 minutes
9. The Eppendorf was placed into the water bath at 42^0C for 45 seconds to allow transformation
10. The Eppendorf was then put back onto ice for 2 minutes, to reduce damage to the bacteria
11. 250 µl SOC medium (without antibiotics) was added
12. The tubes were incubated at 37^0C for one hour while shaking using a shaking incubator
13. Approximately 200 µl of the culture was spread onto LB plates with the appropriate antibiotic. It was grown overnight at 37^0C in an incubator
14. After 12-16 hours, one clone was picked up using a pipette tip. The tip was placed into a 25 ml Falcon tube with 5 ml LB + antibiotic
15. The Falcon tube was incubated at 37^0C overnight in an incubator prior to DNA extraction

2.5.5 Plasmid Extraction from Transformed Bacteria

Plasmid extraction from bacteria was achieved using the MiniPrep Kit (Qiagen) following the manufacturer's protocol. Briefly, the bacterial culture was pelleted by centrifugation at >8000 rpm for 3 minutes at room temperature. The pellet was re-suspended in 250 µl Buffer P1 and transferred to a microcentrifuge tube. 250 µl Buffer P2 was added, and the solution mixed by inverting the tube 6 times. 350 µl Buffer N3 was added, and the solution mixed by inverting the tube 6 times. The solution was centrifuged at 13000 rpm for 10 minutes. The supernatant was added to a QIAprep spin column and centrifuged at 13000 rpm for 1 minute. The flow-through was discarded. 500 µl Buffer PB was applied to the membrane and the tube was spun at 13000 rpm for 1 minute. The flow-through was discarded. 750 µl Buffer PE was added and the tube was spun at 13000 rpm for 1 minute. The flow-through was discarded. The tube was spun again at 13000 rpm for 1 minute, and the flow-through discarded. The QIAprep column was placed into a clean microcentrifuge tube. To elute the DNA, 50 µl Buffer EB was added. The tube was left to stand for 1 minute and was then spun for 1 minute at 13000 rpm.

2.5.5 Sequencing of Plasmid DNA

Sequencing of plasmid DNA was carried out by Dundee Sequencing (https://www.dnaseq.co.uk/). 600 ng of DNA in 20 µl double distilled water was submitted to the laboratory, which use plasmid-appropriate sequencing primers.

2.5.6 Site Directed Mutagenesis (SDM) and Transformation of XL-1 Blue Competent Cells

Site directed mutagenesis was performed using the QuikChange II Kit (Agilent). All primers were designed using Agilent's online primer-design tool (www.agilent.com/genomics/qcpd). Briefly, the protocol involved carrying out a mutant-strand synthesis reaction, using the following reagents:

10X Reaction Buffer	5 µl
dsDNA Template	5-50 ng
125 ng Forward primer	Variable
125 ng Reverse primer	Variable

dNTP mix	1 µl
Water	Variable to a final volume of 50 µl

1 µl of PfuUltra HF DNA polymerase (2.5 U/µl) was added and the reaction mix underwent thermal cycling using the conditions outlined as follows:

95^{0}C	30 seconds
16 cycles of:	
95^{0}C	30 seconds
55^{0}C	1 minute
68^{0}C	7 minutes

After thermal cycling, the reaction mixes were placed on ice for 2 minutes.
1 µl of Dpn1 restriction enzyme was added.

The mutant dsDNA was subsequently used to transform XL-1 Blue Competent cells using the following protocol:

1. XL1-Blue cells were thawed on ice. 50 µl of cells were transferred to a pre-chilled 14 ml BD Falcon polypropylene round-bottom tube
2. 1 µl of the SDM reaction mixture was added to the cells
3. The cells underwent a heat pulse for 45 seconds at 42^{0}C in the waterbath and were then placed on ice for 2 minutes
4. SOC media was preheated to 42^{0}C in the waterbath, and 0.5 ml was added to the cells. The cells were incubated at 37^{0}C for 1 hour whilst shaken at 225-250 rpm in the shaking incubator
5. 250 µl of each transformation reaction was then plated onto ampicillin-containing agar plates
6. The transformation plates were incubated at 37^{0}C for at least 16 hours in an incubator

Following the transformation reactions, colonies were selected and grown as described in section 2.5.4, steps 14 and 15. DNA was extracted using the MiniPrep Kit (2.5.5).

2.5.7 Defrosting HEK293 TCF-Luc Cells

HEK293 TCF-Luc cells were kept in 1 ml aliquots at -80^{0}C .

They were defrosted following the protocol detailed below. Where possible, all steps were performed using a sterile technique in a TC hood:

1. DMEM (1X)-GlutaMax-10% FBS media was warmed to 37^{0}C in the waterbath
2. The HEK293 TCF-Luc cells were thawed at 37^{0}C in the waterbath for 1-3 minutes
3. 9 ml of pre-warmed media and 1ml cells were pipetted into a 15 ml Falcon tube
4. Mixing was carried out by gentle inversion
5. The cells underwent centrifugation for 5 minutes at 250 x g
6. The supernatant was discarded
7. The pellet was re-suspended with 10 ml media
8. The cells underwent centrifugation for 5 minutes at 250 x g
9. The supernatant was discarded
10. The pellet was re-suspended in 5 ml media
11. The suspension was transferred to a 25 cm^2 flask
12. It was kept at 37^{0}C in an incubator

2.5.8 Splitting Cells

Cells were kept in an incubator at 37^{0}C with 5% CO_2. Once they reached approximately 70% confluence, they were split as described below. Where possible, all steps were performed using a sterile technique in the TC hood. Prior to the reactions, DMEM-Glutamax-10% FBS, PBS and Trypsin were pre-warmed to 37^{0}C in the waterbath.

1. The media was removed from the flask
2. 10 ml PBS was added, and the flask was gently agitated to wash the cells
3. The PBS was removed
4. 3 ml 0.05% Trypsin was added
5. The flask was incubated at 37^{0}C for 1 minute in the incubator

6. 7 ml DMEM-Glutamax-10% FBS was added
7. The solution was transferred to a 15 ml Falcon tube
8. The Falcon tube was centrifuged for 5 minutes at 250 x g to create a pellet
9. The cells were re-suspended in 10 ml DMEM-Gluatamax-10% FBS
10. To maintain cells, they were diluted an appropriate amount (e.g. 1:4, 1:6, 1:10 depending on when they were required for the next reaction. Cells were diluted in DMEM-Glutamax-10% FBS) and kept in a 25 cm^2 flask
11. The flask was incubated at 37^0C in an incubator

2.5.9 Plating the Cells for Transfection

Once the cells reached approximately 70% confluence, they were plated for transfection. Steps 1-6 from Section 2.5.8 were followed. Once the DMEM-Glutamax-10% FBS had been added to the cells, they were counted.
20 µl cells were mixed with 20 µl Trypan Blue. 20 µl of the solution was pipetted onto a haemocytometer, and the cells were counted using a light microscope (Leica).

For a full 96-well assay plate, 1 500 000 cells were pipetted from the solution and added to a 15 ml Falcon tube. 10 ml DMEM-Glutamax-10% FBS was added, and the solution was centrifuged for 5 minutes at 250 x g. Following centrifugation, the supernatant was discarded. 10 ml DMEM-Glutamax-10% FBS was added to re-suspend the cells. 100 µl of cell solution (15 000 cells) was added to each well of a 96-well Assay plate. The plate was stored at 37^0C in an incubator.

2.5.10 Transfection

Once the cells were 70-80% confluent they were transfected.

For the Luciferase assays, transfection was performed with the following plasmids:

WT-*AXIN2*, Fearon-*AXIN2* (c.1989G>A), Short-*AXIN2* (c.1642G>T), WT/Short-*AXIN2*. The WT-*AXIN2* and Fearon-*AXIN2* had kindly been supplied by Professor Eric Fearon, University of Michigan, and were N-terminal 6x*myc*-tagged *AXIN2* in

pCS2+MT. Short-*AXIN2* was generated by SDM of the WT-*AXIN2* as described in 2.5.6. For the WT/ Short mix, WT and Short plasmids were used at a ratio of 1:1.

To carry out the reactions, Lipofectamine was diluted in Optimem and incubated for 5 minutes at room temperature. For each well of 15 000 cells, the following volumes were required:

Lipofectamine	0.3 µl
Optimem	24.7µl

A total of 100 ng/well of plasmid was used for transfection.
For each *AXIN2* plasmid, a series of 3-fold dilutions were made, with a maximum amount of 90 ng plasmid (90 ng, 30 ng, 10 ng, 3.33 ng, 1.11 ng, 0.37 ng, 0.12 ng, 0.04 ng). A control with no *AXIN2*-plasmid was also employed. To make the total plasmid mass 100 ng, pcDNA was added to *AXIN2* plasmid.
The reaction volume for each well was made up to 25 µl with Optimem:

AXIN2-Plasmid	Variable: 90-0 ng
pcDNA	Variable: 10-100 ng
Total plasmid:	100ng
Optimem	Variable to make volume up to 25 µl

The Lipofectamine solution was added to the plasmid solution and incubated at room temperature for 20 minutes. The entire reaction mix (50 µl) was subsequently added to a well of cells, and incubated overnight at 37^0C.

For the Luciferase assays, each of the transfection reactions underwent 3 biological replicates, separated in time, and each of the biological replicates underwent three technical replicates. Every experiment included control samples: cells were transfected with empty plasmid, and subsequently did or did not undergo Wnt3a/ R-Spondin stimulation (2.5.11).

2.5.11 Stimulation with Recombinant Human Wnt-3a and Human R-Spondin-1

Following transfection, the cells were stimulated with Recombinant Human Wnt3a and Human R-Spondin-1. 100 µl media was initially removed from each well. 50 µl DMEM-Glutamax-10% FBS was added, containing 20ng/ml R-Spondin and 200 ng/ml Wnt3a.

The cells were incubated at 37^0C overnight.

2.5.12 WST and Luciferase Assays

10 µl WST-1 was added to each well, and the cells were incubated at 37^0C for one hour. WST-1 activity was measured using the Fluorostar Optima. Readings were taken at 450 nm and 590 nm.

The assay plate was subsequently inverted, and all liquid tapped out. 35 µl Glo-Lysis buffer was added to each well, and the assay plate was mixed using the Stuart Gyro Rocker for 15 minutes at 45 rpm. 35 µl Glo-Substrate was then added to each well. Luciferase activity was read using the Fluorostar Optima. Luciferase activity was normalised to the WST-Score (Luciferase score/ (WST 450/WST 590)).

The normalised luciferase activity in the cells which had been transfected with empty vector and which had been stimulated with R-Spondin/ Wnt3a was given a score of 1. The normalised luciferase activity in the cells transfected with *AXIN2* plasmid was compared to this.

2.5.13 Cell Lysis and Protein Extraction

The cell culture dish was placed on ice and the media removed. Cells were washed with ice cold PBS, which was then removed. 150 µl of cell lysis buffer was added to each well. 10 ml buffer was composed of:

NaCl (5 M)	0.3 ml
Tris (1 M)	0.5 ml
IGEPAL CA630	0.05 ml

Roche Complete protease inhibitor (1 tablet per 10 ml)
Roche PhosphoSTOP phosphatase inhibitor (1 tablet per 50 ml)
Glycerol (10%) 1 ml
Water 8.15 ml

Adherent cells were scraped off the dish using a cold cell scraper. The cell suspension was transferred to a pre-cooled microcentrifuge tube and was agitated at 4^0C for 30 minutes. The microcentrifuge tubes were subsequently centrifuged at 4^0C for 30 minutes at 13 000 rpm. The tubes were then placed on ice. The supernatant was aspirated and placed into a fresh microcentrifuge tube. Protein-containing lysate was stored at -80^0C.

2.5.14 Protein Quantification using a Bicinchoninic Acid (BCA) Assay

A BSA protein standard was prepared according to the following volumes:

Standard (µg/ml)	**µl of 10 mg/ml BSA**	**µl H_2O**
500	25	475
400	20	480
300	15	485
200	10	490
100	5	495
50	2.5	497.5
0	0	500

Table 2.5: BSA protein standards

50 µl of each standard was added in duplicate to a 96-well plate.

1 µl of protein lysate was added to 199 µl water, to make a 1:200 dilution. 50 µl of this was added in duplicate to the 96-well plate. The stock solution was then diluted by a factor of 2 to give a 1:400 dilution. 50 µl of this was added in duplicate to the 96-well plate.

The BCA reagent was prepared by mixing components A and B at a ratio of 50:1. 75 µl of BCA reagent was added to each well of standard or sample. The plate was incubated at room temperature for 30 minutes with gentle agitation.

The absorbance was read at 590 nm using the Fluorostar Optima, and a standard curve was used to calculate the protein content in each sample.

2.5.15 Western Blotting

677 µg of each protein in 15 µl water was mixed with 5 µl NuPAGE LDS Sample Buffer. Samples were heated for 5 minutes at 95^0C using the QBD2 heating block. They were loaded onto a NuPAGE 4-12% Bis Tris Minigel. The protein marker used was the Novex Sharp Pre-Stained Protein Standard. The samples were electrophoresed for 1.5 hours at 200 V using NuPAGE MES SDS running buffer.

The protein was transferred from the gel to a nitrocellulose membrane using the iblot system (Invitrogen).

The membrane was blocked with 5% non-fat powdered milk in TBS for one hour at room temperature, with constant agitation. The membranes were then incubated with anti-myc-tag primary monoclonal antibody (9B11 New England Biolabs) diluted in TBS-Milk overnight at 4^0C with constant agitation.

The membrane was washed twice in water, and was incubated with anti-mouse secondary antibody (NXA931 Sigma Aldrich) diluted in TBS-milk, for one hour at room temperature with constant agitation.

The membrane was washed twice in water. It was incubated for 5 minutes at room temperature with 8 ml Supersignal West Pico Chemiluminescent Substrate.

Antibody detection was achieved using the BioRad ChemiDoc MP Imaging System.

2.5.16 Stripping Nitrocellulose Membranes and β-Actin Detection

The nitrocellulose membrane was washed with TBS. It was immersed in Restore Plus Western Blot Stripping Buffer and incubated for 15 minutes at room temperature. The buffer was removed, and the membrane washed with TBS.

The membrane was blocked with 5% non-fat powdered milk in TBS for one hour at room temperature, with constant agitation. To detect β-actin, the primary antibody used was A5441 (Sigma Aldrich), and the further steps followed are those described above (2.5.15).

Chapter 3 *APC* and *MUTYH* in Colorectal Polyposis

3.1 Introduction

The aim of this chapter was to identify missed or novel genetic variants in *APC* or *MUTYH* responsible for the polyposis phenotype in NMI patients. Such variants may occur outside of the open reading frame (ORF) of the genes, in promoters, untranslated regions (UTRs) or introns, or they may be conventional mutations which occur at a low frequency within the patients (mosaicism). They could underpin the pathogenesis of colorectal neoplasia through mechanisms such as abnormal gene transcription, reduced stability of mRNA and abnormal mRNA splicing. I sought to identify variants and their mechanisms of action through a combination of approaches (Figure 3.1):

- Quantitative PCR (qPCR) to interrogate *APC* and *MUTYH* transcription
- gDNA/ cDNA sequencing to screen for allelic imbalance (AI)
- cDNA sequencing of the coding sequences of *APC* and *MUTYH* to screen for splicing abnormalities
- Ultradeep sequencing (UDS) of the whole of the *APC* and *MUTYH* genomic loci to identify variants outside of the ORF and low frequency variants, i.e. patients with generalised mosaicism

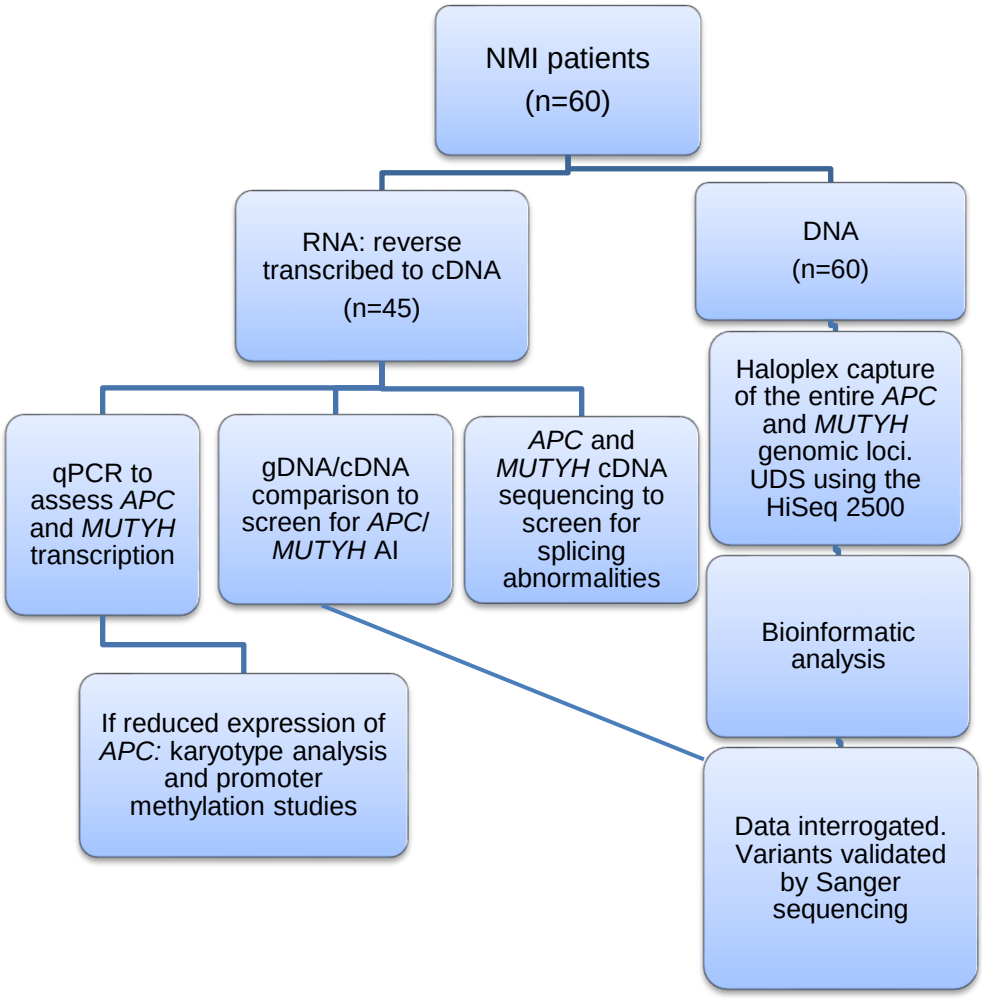


Figure 3.1 Flow chart of the studies employed in the interrogation of APC and MUTYH

3.2 Methods

3.2.1 Quantitative PCR (qPCR): *APC* and *MUTYH* Transcription

To determine whether any NMI polyposis patient had evidence of reduced transcription of *APC* or *MUTYH* compared to a cohort of healthy controls, RNA levels within a control cohort needed to be determined. To date, no study has characterised the normal range of *APC* and *MUTYH* RNA levels in RNA derived from venous blood samples from a healthy control cohort.

High quality RNA was available for 40 healthy control samples (2.3.16) and 45 NMI patients (2.2.3.4). 3 affected FAP control patients and 4 affected MAP control patients were also included (Table 3.1). FAPPol51 was used as a positive control for reduced *APC* transcription due to the confirmed presence by the NHS diagnostic service of an *APC* promoter deletion known to impair transcription. The remaining FAP and MAP patients have mutations which are not predicted to affect transcription, so were used as negative controls. Although some of these mutations may result in nonsense mediated decay (NMD) or could potentially affect transcript stability, the Taqman probes generate such short amplicons that it is highly likely they will be able to bind to fragmented RNA/ cDNA so that levels will appear to be within the normal range. Unfortunately, *MUTYH* promoter mutations resulting in reduced gene transcription are thus far unreported in the literature, so no positive controls for reduced *MUTYH* transcription were available.

Sample	**Sample Group**	**Mutation**	**Expected Result for *APC* and *MUTYH* Gene Transcription**
Halo37	MAP	Compound heterozygous for *MUTYH* c.536A>G & c.649C>T	Normal
Halo59	MAP	*MUTYH* c.1187G>A homozygote	Normal
MAPPol71	MAP	Compound heterozygous for *MUTYH* c.303G>T and c.312C>A	Normal
MAPPol90	MAP	Compound heterozygous for *MUTYH* c.1187G>A and c.536A>G	Normal
FAPPol51	FAP	*APC* promoter 1B deletion	Reduced *APC* transcription
FAPPol141	FAP	*APC* exons 11 and 12 deleted	Normal
Halo42	FAP	*APC* c.1187 dup A, *MUTYH* c.536A>G	Normal

Table 3.1: FAP and MAP controls used in gene transcription studies

RNA was converted to cDNA following standard protocols (2.3.3.2). cDNA underwent qPCR using standard reagents and reaction conditions (2.3.15) and Taqman assays for *ACTB, GAPDH, APC* and *MUTYH* (2.3.16.1, Table 2.3). Results were analysed using ThermoFisher Cloud software (2.3.16.1.1). Any patient with an Rq </= 0.6 when compared to the healthy control cohort was identified for further characterisation.

3.2.2 Karyotype Analysis

Chromosomal rearrangements can affect gene expression through mechanisms such as disrupting genes or regulatory elements, producing fusion genes, placing genes in anomalous chromatin environments, placing genes near to telomeres and aberrant nuclear positioning (Harewood and Fraser 2014). Karyotype analysis was performed by the AWMGS following their Standard Operating Procedure (SOP).

3.2.3 Promoter Methylation Studies

MSP was carried out to determine the methylation status of the *APC* 1A and 1B promoters (2.3.14). MSP for the 1A promoter used primers described by Arnold *et al* (2004) (Appendix 3.1). MSP for the 1B promoter used primers desribed by Esteller *et al* (2000) (Appendix 3.2).

3.2.4 RNA Studies: Allelic Imbalance (AI) and Splicing Abnormalities

45 NMI patients and 4 unaffected relatives (Halo75, Halo76, Halo77 and Halo78) underwent studies on DNA and RNA to screen for *APC* +/- *MUTYH* RNA AI. All of the work done with *MUTYH* was carried out by myself. I carried out the *APC* AI work on 22 samples. The remaining work was completed by Alice Davies and Alice Bolton, intercalating medical students (Davies A 2015, Bolton A 2016), and Julie Maynard, research assistant, Cardiff University.

To screen for AI, common non-pathogenic exonic SNPs were identified in *APC* and *MUTYH*, using UCSC Genome Bioinformatics (http://genome.ucsc.edu/cgi-bin/hgTracks?db=hg19&position=chr1%3A45794914-45806142&hgsid=414232457_bzuux4uOZeA1wBLT5vBhLtb6YSg9) and NCBI dbSNP software (http://www.ncbi.nlm.nih.gov/projects/SNP/) (accessed 26/11/2014). The *APC* SNPs were rs2229992, rs351771, rs41115, rs42427, rs866006 and rs465899. The *MUTYH* SNPs were rs3219489 and rs3219497 (full details in Appendix 3.3).

Comparison was made of the sequence at each of these SNPs in the patient's gDNA and their cDNA. If a patient was heterozygous for a common SNP in the

germline but was found to be hemizygous for either allele in cDNA, this was suggestive of allelic imbalance.

3.2.4.1 gDNA analysis

gDNA analysis was initially carried out by examining the UDS results (3.2.5). Any heterozygous SNPs identified were validated with Sanger Sequencing. One of the SNPs, rs465899, was not covered adequately through UDS, therefore Sanger sequencing was performed on all patients for this locus. Sanger sequencing and sequence analysis were undertaken using standard reagents and reaction conditions (2.3.8.1, 2.3.8.5, 2.3.9-2.3.13). Primers were those used by the AWMGS (Appendix 3.4). Two samples required optimisation (Appendix 3.5).

3.2.4.2 RNA Analysis

RNA analysis involved reverse transcription to cDNA (2.3.3.2), PCR with gene specific primers (Appendix 3.6), and sequence analysis. Standard protocols and reaction conditions were used (2.3.8.1, 2.3.8.5, 2.3.9-2.3.13). Only those SNPs which were confirmed to be heterozygous in the germline were examined in the RNA, as homozygous SNPs do not yield any discriminating information. PCR conditions were optimised as shown in Appendices 3.7 (*APC*) and 3.8 (*MUTYH*).

3.2.4.3 *APC* and *MUTYH* Splicing Abnormalities

45 NMI patients underwent sequencing of *APC* +/- *MUTYH* cDNA to screen for splicing abnormalities. All *MUTYH* sequencing was carried out by myself. I carried out the *APC* cDNA sequencing for 17 samples. The remaining work was completed by Alice Davies and Alice Bolton, intercalating medical students (Davies A 2015, Bolton A 2016), and Julie Maynard, research assistant, Cardiff University.

3.2.4.3.1 *APC* cDNA Sequencing

RNA was reverse transcribed (2.3.3.2) and cDNA underwent PCR with gene specific primers (Appendix 3.9). Standard reagents and reaction conditions were

employed for PCR and sequencing (2.3.8.1, 2.3.8.5, 2.3.9-2.3.13). The details for the samples which required optimisation are in Appendix 3.10.

3.2.4.3.2 *MUTYH* cDNA Sequencing

MAP is an autosomal recessive (AR) disease. Even if abnormal splicing is found in one allele, there would need to be a pathogenic variant present in the other allele for manifestation of polyposis. In view of this, a small study was initially performed on 9 NMI polyposis patients, in order to optimise reaction conditions. These were the first study participants that RNA was available for. Subsequently, *MUTYH* cDNA sequencing was completed on all those patients who were known to have/ potentially have a pathogenic variant present affecting *MUTYH*:

- Known monoallelic carriers of a *MUTYH* mutation
    - Halo40
    - Halo41
- Potentially reduced expression of *MUTYH* (3.3.1)
    - Halo36
    - Halo63
    - Halo71

RNA was reverse transcribed (2.3.3.2) and cDNA underwent PCR with gene specific primers (Appendix 3.11). Standard reagents and reaction conditions were employed for PCR and sequencing (2.3.8.1, 2.3.8.5, 2.3.9-2.3.13). The details for the samples which required optimisation are in Appendix 3.12.

3.2.5 *APC* and *MUTYH* Capture and Ultradeep Sequencing (UDS)

UDS was performed for the whole genomic loci of *APC* and *MUTYH* (chr5:112042936-112186350, chr1:45794202-45807013) to identify mutations outside of the ORF and to identify *APC* mutations which occur at a low frequency within an individual. It was carried out on DNA extracted from whole blood (2.3.2). Target sequence capture was achieved using the Haloplex assay (2.4.1).

Sequencing was performed by the Wales Gene Park (WGP) using the HiSeq 2500 (2.4.2).
82 individuals underwent UDS:

- 60 NMI polyposis patients (Table 2.1)
- 11 healthy controls (Tables 2.1 and 2.2): Halo02, Halo03, Halo09, Halo11, Halo73, Halo74, Halo75, Halo76, Halo77, Halo78, and Halo95
- 11 positive control samples (Table 2.2): 8 FAP patients (Halo01, F014_M_004_1, F014_M_005_1, F014_M_006_1, Halo10, Halo12, Halo16, Halo42 (also carries a monoallelic *MUTYH* mutation) and 3 MAP patients (Halo37, Halo39, Halo59)

Bioinformatic analysis was performed by Dr. Peter Giles, WGP (2.4.3).

3.2.5.1 Variant Analysis and Selection

Variants of interest were selected using the following criteria:

- Rare variants, with an allele frequency </= 1% according to dbSNP data or The 1000 Genomes Project data
- Variants which were present in all affected family members
- Synonymous variants were excluded
- Variants which had been identified previously by the AWMGS diagnostic service and which were known to be non-pathogenic were excluded
- Variants found in unaffected relatives/ healthy controls were excluded, taking into account the mode of inheritance (i.e. if a patient was found to be homozygous for a *MUTYH* variant, it was only excluded if it was also homozygous in an unaffected relative/ healthy control).
- Exonic missense variants present in <2% of reads which were identified in multiple samples were excluded as they were likely to represent sequencing artefacts

All shortlisted variants were then analysed using CADD software (http://cadd.gs.washington.edu/home) and were assessed using the Integrative Genomics Viewer (IGV) (https://www.broadinstitute.org/igv/). Variants which had a CADD score >/= 15 were validated with Sanger sequencing.

3.2.5.2 Validation of Identified Variants

Variants were validated with Sanger Sequencing, using standard reagents and reaction conditions (2.3.8.1, 2.3.8.5, 2.3.9-2.3.13). Primers are listed in Appendix 3.13. Variants present at a low frequency (<20%) were validated with COLD-PCR, using the *full-* or *fast-* standard protocol depending on the nature of the variant (2.3.8.3, 2.3.8.4, 2.3.8.6, 2.3.8.7).

3.3 Results

3.3.1 Quantitative PCR (qPCR): *APC* and *MUTYH* Expression

A total of 92 samples underwent successful qPCR to determine levels of *APC* and *MUTYH* transcription: 40 healthy controls, 45 NMI polyposis patients, 3 FAP controls and 4 MAP controls (2.3.16, 2.2.3.5, Table 3.1). One of the FAP controls, FAPPol51, had an *APC* promoter deletion known to impair *APC* transcription. The sample was used as a positive control to ensure that the assay was capable of detecting reduced *APC* transcription. The remaining FAP and MAP controls all had genetic variants which did not affect gene transcription: they were included as negative controls to confirm that the assay identified that *APC* and *MUTYH* RNA levels were within the normal range.

The results for the control samples were largely as anticipated, although one of the FAP controls who also carries a monoallelic *MUTYH* mutation, Halo42, appeared to have reduced *MUTYH* transcription, although this was not statistically significant (Table 3.2).

Sample	Sample Group	Mutation	Expected Result	*APC* Rq Value	*MUTYH* Rq Value
Halo37	MAP	Compound heterozygous for *MUTYH* c.536A>G & c.649C>T	Normal	0.759	0.848
Halo59	MAP	*MUTYH* c.1187G>A homozygote	Normal	1.102	0.896
MAPPol71	MAP	Compound heterozygous for *MUTYH* c.303G>T and c.312C>A	Normal	1.081	0.963
MAPPol90	MAP	Compound heterozygous for *MUTYH* c.1187G>A and c.536A>G	Normal	0.796	0.779
FAPPol51	FAP	*APC* promoter 1B deletion	Reduced *APC* transcription	0.484 (p=0.03)	0.903
FAPPol141	FAP	*APC* exons 11 and 12 deleted	Normal	0.772	0.944
Halo42	FAP	*APC* c.1187dupA, *MUTYH* c.536A>G	Normal	0.773	0.560 (p=0.15)

Table 3.2: Control results for APC and MUTYH transcription studies

The individual results for all the NMI patients are in Appendix 3.14 .

Four patients were found to have *APC* Rq values below 0.6: Halo46, *APC* Rq= 0.563 (p=0.09); Halo52, *APC* Rq=0.368 (p=0.003); Halo53, *APC* Rq=0.545 (p=0.07); Halo64, *APC* Rq=0.417 (p=0.01). These results were confirmed by repeating the experiment from the initial reverse transcription reaction (Table 3.3).

Three patients had *MUTYH* Rq values below 0.6. Although these results are interesting, they do not have the same clinical implications as do the *APC* gene transcription results. MAP is an AR disease, therefore reduced transcription of one allele alone would not explain a clinical phenotype. In addition to this, a negative control was also observed to have an apparently reduced *MUTYH* Rq value (although this was not statistically significant), a positive control was not available, and only the Rq value for Halo36 reached statistical significance, with p=0.03. The results were not repeated, but these patients were selected for further RNA studies (cDNA sequencing, 3.2.4.3.2) to determine whether they had any further potentially disease-causing genetic aberrations.

The phenotype of those patients with an Rq value </= 0.6 is summarised in Table 3.3:

Patient Identification	***APC/ MUTYH* Rq Value**	**Repeat Result**	**Clinical Phenotype**
Halo46	*APC* Rq = 0.563	*	44-year-old female who clinically has FAP and who has undergone a colectomy and proctectomy
Halo52	*APC* Rq = 0.368	*APC* Rq = 0.400	59-year-old male with >1000 adenomas
Halo53	*APC* Rq = 0.545	*APC* Rq = 0.527	51-year-old male with approximately 400 adenomas and CRC at age 27
Halo64	*APC* Rq = 0.417	*APC* Rq = 0.291	54-year-old female with thousands of colorectal adenomas and a caecal carcinoma at age 23
Halo36	*MUTYH* Rq = 0.41		67-year-old male with >10 adenomas
Halo63	*MUTYH* Rq = 0.509		68-year-old male with at least 24 polyps 5 biopsies show 4 adenomas and 1 HPP
Halo71	*MUTYH* Rq = 0.508		65-year-old female with 14 adenomas

Table 3.3: Phenotypes of patients with evidence of reduced transcription of APC/ MUTYH

*Repeat testing was not carried out for Halo46, as the causative mutation was identified (3.3.5.3).

3.3.2 Karyotype Analysis

Karyotype analysis was performed on Halo52 and Halo53, both of whom appeared to have evidence of reduced *APC* expression. Both had normal chromosome complements. Unfortunately, patient Halo64 has been lost to follow up, so it was not possible to obtain a fresh blood sample. Halo46 did not undergo karyotyping as a pathogenic mutation was identified (see section 3.3.5.3).

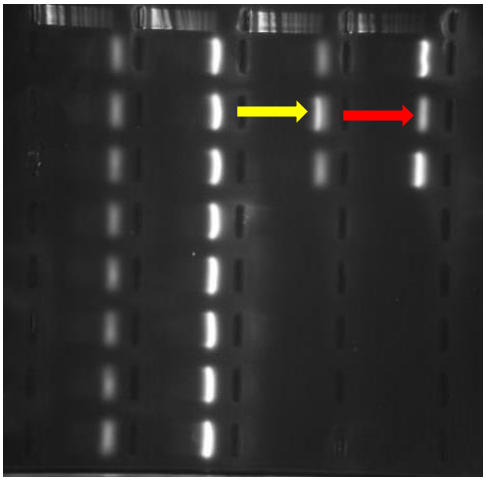
3.3.3 Promoter Methylation Studies

Methylation analysis of the *APC* 1A and 1B promoters was performed on 3 patients with reduced *APC* transcription (Halo52, Halo53 and Halo64) and the unaffected mother of Halo53. It was not done for Halo46 as the pathogenic mutation was identified (section 3.3.5.3). Methylation studies were also performed for NMI patients with any evidence of *APC* AI (section 3.3.4.1): Halo08 and her affected son, Halo51 and her unaffected parents.

There was no evidence of promoter methylation for either promoter:

M primers U primers M primers U primers

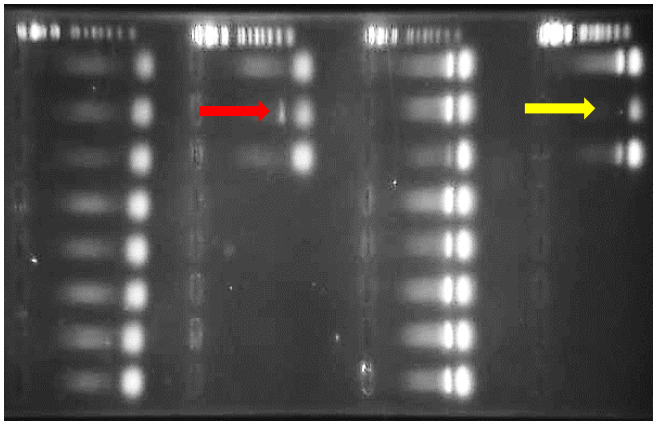
Ladder
1. Halo53
2. Halo53 Mother
3. Halo52
4. Halo64
5. Halo51
6. Halo51 Mother
7. Halo51 Father
8. Halo08



Ladder
9. Halo08 Son
10. Control M DNA
11. Control U DNA

Figure 3.2. MSP for APC promoter 1A. All samples, apart from the control methylated DNA (M DNA: red arrow), produce a PCR product following MSP using primers specific for unmethylated DNA (U primers). This is demonstrated by the high intensity of the bands. The only PCR product obtained when MSP was undertaken using primers specific for methylated-DNA (M primers) was that from the control methylated DNA sample (M DNA: yellow arrow), as shown by the high intensity band.

M primers M primers U primers U primers



Ladder
1. Halo53
2. Halo53 Mother
3. Halo52
4. Halo64
5. Halo51
6. Halo51 Mother
7. Halo51 Father
8. Halo08

Ladder
9. Halo08 Son
10. Control M DNA
11. Control U DNA

Figure 3.3 MSP for APC promoter 1B. All samples, apart from the control methylated DNA (M DNA: yellow arrow), produce a PCR product following MSP using primers specific for unmethylated DNA (U primers) (there was a product for Halo52, although the PCR band was fainter than those obtained for the other samples). The only PCR product obtained when MSP was undertaken using primers specific for methylated DNA (M primers) was that from the control methylated DNA sample (M DNA: red arrow). The bands which are smaller than the PCR products represent primer-dimers.

3.3.4 Allelic Imbalance Studies

Allelic Imbalance (AI) studies were completed on 45 NMI patients and 4 unaffected relatives for whom high quality RNA samples were available (Halo75, Halo76, Halo77, Halo78).

3.3.4.1 *APC* Allelic Imbalance

Of the 45 patient samples, 22 had informative results, i.e. the presence of at least one heterozygous SNP in the germline. A total of 4 patients displayed AI: 18% of the informative cohort, or 9% overall. The patients were Halo08 and Halo51 (both had AI at the final SNP and visual inspection of the sequencing traces were suggestive of one allele having a weaker signal at some of the preceding SNPs), Halo52 (AI at first SNP only. The subsequent SNPs were uninformative) and Halo53 (AI throughout). All of the individual patient results are in Appendix 3.15. The sequencing results for the SNPs exhibiting AI are in Appendix 3.16.

Patients Halo52 and Halo53 had results which would be consistent with absent expression from one allele, which may result from a promoter lesion (Charames *et al* 2008). This was concordant with their gene expression results (3.3.1). It is noted that patient Halo53 did display tiny peaks of the alternate allele at SNPs rs41115 and rs866006 (Appendix 3.16). Whilst this may represent a sequencing artefact, it is possible that this allele was being expressed at such low levels that it was not reliably detected with Sanger sequencing.

Patients Halo08 and Halo51 displayed complete loss of signal for one allele at the final SNP only. This could possibly result from allele specific transcription termination or mRNA degradation, the presence of an allele specific alternate transcription start site (Wagner *et al* 2010) or it could represent a sequencing artefact. There also appeared to be an abnormal allelic expression ratio at some of the preceding SNPs. These findings were confirmed by repeating the experiments (Appendix 3.17). The significance of these findings was not confirmed.

The parents of Halo51 and Halo53 did not exhibit AI.

3.3.4.2 *MUTYH* Allelic Imbalance

45 NMI polyposis patients underwent *MUTYH* AI studies. 16 had informative results. There was no evidence of AI in any sample. The full results are in Appendix 3.18.

3.3.5 *APC* and *MUTYH* Splicing Abnormalities

APC cDNA sequencing was completed on 45 patients. There was no evidence of any splicing abnormalities.

MUTYH cDNA sequencing was completed on 14 patients, including those with a monoallelic *MUTYH* mutation and those with potentially reduced *MUTYH* expression. There was no evidence of any splicing abnormalities.

3.3.5 *APC* and *MUTYH* Capture and Ultradeep Sequencing (UDS)

The Haloplex assay was used for targeted capture of the whole of the *APC* and *MUTYH* genomic loci, followed by UDS using the HiSeq 2500 (Illumina) (Three samples had undergone sequencing using the MiSeq (Illumina) before the HiSeq 2500 was being utilised: F014_M_004_1, F014_M_005_1, F014_M_006_1).

3.3.5.1 Coverage of Ultradeep Sequencing

The average depth of coverage of *APC* and *MUTYH* using the HiSeq 2500 was 1665.5x and 2575.2x respectively. The mean on-target coverage for *APC* was 97.2% and for *MUTYH* it was 97.3%. For *APC*, regions with consistently low coverage (<50 reads) occurred in exons 3 (65 bp), 5 (65 bp, 14bp), 6 (27 bp) and 15 (9 bp, 11 bp, 15 bp, 14 bp, 65 bp, 23 bp). There were 2 small areas with consistently no coverage: 16 bp of exon 5 and 52 bp of exon 15. For *MUTYH* there were 9 bp of exon 12 with low coverage. The detailed coverage for each individual patient is recorded in Appendices 3.19 (*APC*) and 3.20 (*MUTYH*).

3.3.5.2 Control Results

The Haloplex assay followed by UDS identified all 15 known variants in *APC* or *MUTYH*. This included 8 FAP controls, one of whom also carried a *MUTYH* variant (Halo42), 3 MAP controls, a previously confirmed *APC* variant in Halo33 and previously confirmed *MUTYH* variants in Halo40 and Halo41. These variants and the coverage are described in Table 3.4:

Sample Identification	Gene	Variant	Mean Coverage of Gene with Variant
F014_M_004_1	*APC*	c.2805C>A	2410
F014_M_005_1	*APC*	c.4393-4394delGA	3118
F014_M_006_1	*APC*	5% mosaic *APC* c.4393-4394delGA	2421
Halo01	*APC*	c.3631-3632delAT	723
Halo10	*APC*	c.3183-3187delACAAA	328
Halo12	*APC*	c.3927-3931delAAAGA	182
Halo16	*APC*	c.3408delA	236
Halo33	*APC*	c.6363-6365dupTGC	235
Halo42	*APC*	c.1133dupA	408
Halo37	*MUTYH*	Compound heterozygote: c.536A>G & c.649C>T	3994
Halo39	*MUTYH*	Compound heterozygote: c.303G>T and c.312C>A	638
Halo40	*MUTYH*	c.920G>A heterozygote	2443
Halo41	*MUTYH*	c.1187G>A heterozygote	7636
Halo42	*MUTYH*	c.536A>G heterozygote	604
Halo59	*MUTYH*	c.1187G>A homozygote	3387

Table 3.4: Control samples used for validating Haloplex capture and UDS

3.3.5.3 Patient Results

60 NMI polyposis patients underwent UDS of *APC* and *MUTYH*. 15 variants selected for validation/ further investigation are listed in Appendix 3.21. Variants

present in >20% of reads were validated with standard Sanger sequencing following standard PCR protocols (Appendix 3.22). Low frequency variants, present in <20% of reads, were amplified with COLD-PCR prior to Sanger sequencing (Appendix 3.23). Validation was not carried out if the variant had been previously identified by the AWMGS.

Table 3.5 summarises the variants which were validated.

Sample ID	Gene	Position	Variant	Location	Effect	Additional Information	CADD Score
Halo23	*APC*	5: 112043492	C>A	5'UTR/ Exon 1	N/A	None	21.1
Halo30	*APC*	5: 112102960	C>T	Exonic	Missense	Validated by NHS	34
Halo35	*APC*	5:112104652	T>A	Intronic	N/A	None	18.02
Halo46	*APC*	5: 112043225	G>A	5'UTR	N/A	None	22.4
Halo51	*APC*	5: 112095775	T>A	Intronic	N/A	None	16.36
Halo56	*APC*	5: 112043282	C>CG	5'UTR	N/A	None	16.52
Halo62	*APC*	5: 112162474	C>T	Intron	N/A	None	18.98
Halo66	*APC*	5: 112163697	A>C	Exonic	Missense	Validated by NHS	25.4
Halo71	*APC*	5: 112102960	C>T	Exonic	Missense	Validated by NHS	34
Halo40	*MUTYH*	1: 45797851	C>T	Exonic	Missense	Validated by NHS. Patient only has one pathogenic mutation	20.4

Table 3.5 Validated APC and MUTYH variants

All the variants which appeared to be approximately heterozygous were validated. None of the low frequency variants were definitely real, although one was uncertain, a missense exonic variant in sample Halo45, present in 4% of reads. A truncating *APC* exonic variant in Halo63 was not validated with Sanger sequencing. However, there was low coverage at this locus with NGS, with the variant being present in only 2/11 reads, so it is likely to be an artefact. The validated variants will be appraised in the Chapter Discussion (3.4.3.2).

Table 3.6 summarises the key positive findings from Chapter 3:

Unique Identifier	**qPCR: *APC* Rq**	**Karyotype Analysis**	***APC* Promoter Methylation**	***APC* AI**	***APC* cDNA Sequencing**	***APC* UDS**
Halo46	0.563	N/A	N/A	Uninformative	Normal	Chr5: 112043225 G>A
Halo52	0.368	Normal	Normal	AI at first SNP	Normal	Nil
Halo53	0.545	Normal	Normal	AI throughout	Normal	Nil
Halo64	0.417	Not done	Normal	Uninformative	Normal	Nil

Table 3.6 Key positive results from all APC and MUTYH studies

3.4 Discussion

This study involved investigating patients with multiple colorectal adenomas (>10) who had no mutation identified (NMI) in *APC* or *MUTYH* during routine clinical diagnosis. The aim of this chapter was to identify missed or novel genetic variants in *APC* or *MUTYH* responsible for tumourigenesis. I sought to identify such variants through a combination of:

- Quantitative PCR (qPCR) to interrogate *APC* and *MUTYH* transcription
- gDNA/ cDNA sequencing to screen for allelic imbalance (AI)
- cDNA sequencing of the coding sequences of *APC* and *MUTYH* to screen for splicing abnormalities
- Ultradeep sequencing (UDS) of the whole of the *APC* and *MUTYH* genes to identify variants outside of the ORF and *APC* low frequency variants

DNA samples were available for all 60 patients, and RNA samples were available for 45 NMI polyposis patients.

3.4.1 Gene Expression Studies and Allelic Imbalance

RNA studies yielded interesting results for 7 patients: 4 appeared to have reduced *APC* transcription and 3 appeared to have reduced *MUTYH* transcription. The 4 patients with reduced *APC* transcription all had phenotypes in keeping with a clinical diagnosis of FAP. The gene transcription results were concordant with the AI results in Halo52 and Halo53, while AI studies were uninformative in Halo46 and Halo64.

The cause of the reduced *APC* expression was determined in Halo46. UDS identified a 5'UTR variant (G>A at 5:112043225), at position c.-190G>A. This variant is predicted to be pathogenic, with a CADD score of 22.4. The patient has a strong family history of polyposis/ CRC, with the inheritance pattern suggestive of an autosomal dominant trait:

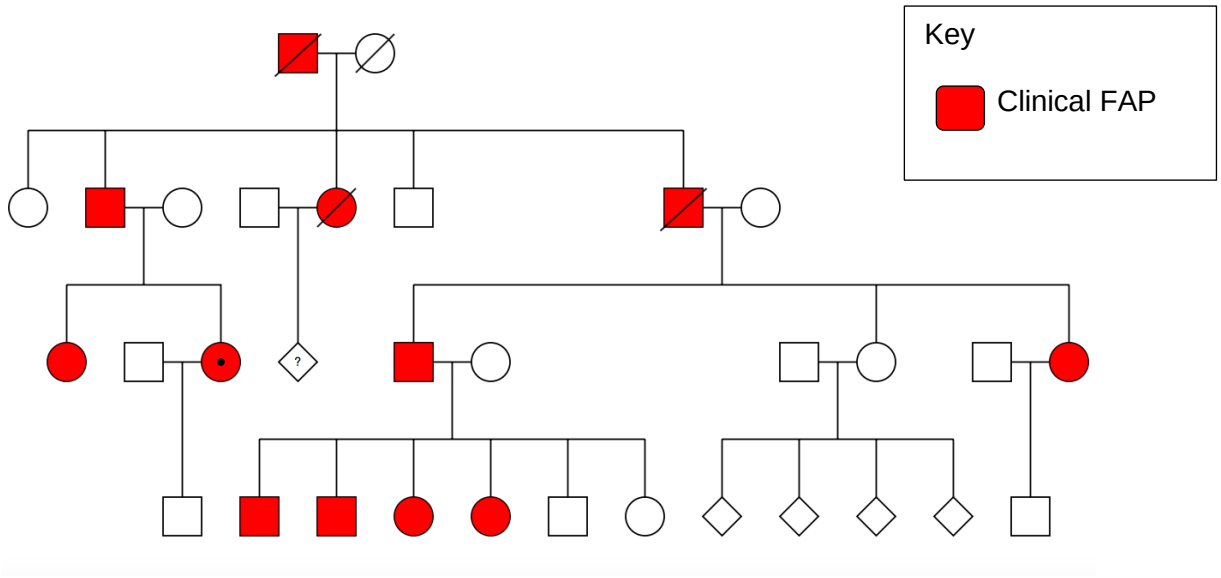


Figure 3.4 Family tree of Halo46. Halo46 is marked with a black circle

At the time this variant was identified in the index case, a paper was published which describes the exact same variant (Li *et al* 2016). The variant occurred in a family with 5 affected patients, over 3 generations, with polyposis and/ or CRC. The group performed electromobility shift assays (EMSA) to show that the mutation led to reduced protein binding, and that the protein likely to be affected was the transcription factor YY1. Further work demonstrated that carriers of the mutation also showed AI of *APC*. The work described by Li's paper supports the pathogenicity of the variant identified in Halo46. Furthermore, we were able to recruit the patient's father and one cousin, both whom have clinical FAP. Both were found to carry the same mutation. qPCR of *APC* cDNA levels was also performed: the Rq value was 0.55 when compared to the healthy control cohort for the father, and 0.63 for the cousin. I tried to recruit further family members to further investigate whether the variant was segregating with the disease phenotype, but unfortunately was unable to do so during the time course of this project. However, this research finding is being translated into clinical practice, and the AWMGS are establishing predictive genetic testing for this specific mutation in members of the extended family. As far as we are aware, Halo46 does not have a gastric phenotype. This is an interesting observation, as point mutations in the *APC* 1B promoter are reported to result in gastric neoplasia (Li *et al* 2016). It is possible that Halo46 actually does have such a phenotype, but that we do not know about - for example if she has not attended upper gastrointestinal (UGI) endoscopic surveillance, or if she has attended UGI endoscopy but the associated endoscopic reports/ histopathological reports were not sent to the AWMGS for inclusion in the

genetics notes. Alternatively, she may lack gastric neoplasia if her gastric mucosa is 'protected' in some way. The *APC* 1B transcript is usually more important than the 1A transcript in the stomach (reviewed in Li *et al* 2016) but it is possible that in this family, the 1A transcript plays a significant role. Gastric protection could also be the result of environmental factors such as the biome. It is also feasible that this specific mutation has a low/ incomplete penetrance in the stomach or has variable expressivity, mediated through the effects of modifier genes.

I sought to determine the cause of the reduced gene transcription in the remaining 3 samples. Karyotype analysis was normal for Halo52 and Halo53, and there was no evidence of promoter methylation for any of the 3 patients. If further evidence regarding the lack of promoter methylation was required, additonal techniques to confirm this would include pyrosequencing or qPCR. UDS results did not identify any variants which could underpin abnormal transcription. The cause of the reduced gene transcription was therefore not apparent. There are several possible explanations for this phenomenon, including upstream/ enhancer variants which were not captured by the Haloplex assay, abnormalities in transcription factor(s), variants in further genes which are modulating gene expression and epigenetic effects. All 3 patients underwent whole exome sequencing (Chapter 4) but further investigation into the reduced gene transcription was beyond the scope of this project. We shall recommend that the patients undergo whole genome sequencing with the aim of searching for variants in the regions of the genome which have not yet been interrogated.

Three patients appeared to have a reduction in *MUTYH* transcription (Halo36, Halo63 and Halo71), although only the result for Halo36 was statistically significant. All AI results were uninformative. However, even if the reduced transcription is real, this alone would not explain the patients' phenotypes, as MAP is an AR disease, and there would have to be a pathogenic change in the second allele: in all three cases *MUTYH* cDNA sequencing showed no splicing abnormalities, and UDS did not reveal any pathogenic *MUTYH* variants. Furthermore, reduced *MUTYH* transcription was also observed in a negative control sample, and it was not possible to identify a positive control, so the results are weak and are unlikely to be clinically significant.

To investigate transcription levels of *APC* and *MUTYH* it was necessary to determine the range of transcription of these genes in a healthy control cohort. Gene transcription studies were carried out on 40 unaffected individuals to generate this data. To improve the reliability of the results, it would be necessary to increase the population size of healthy individuals. If large-scale data were obtained, its validity would be more certain, and this could offer the possibility of gene transcription studies being performed in a diagnostic setting. At the moment, it is cautiously suggested that gene transcription studies could be considered as a screening test in NMI patients, but that they need to be interpreted in the context of additional investigations such as AI studies and whole gene sequencing.

Gene transcription studies were carried out using one probe for each gene. For *APC,* the probe spanned the junction between exons 12 and 13 (HS01568269_m1) and for *MUTYH* the probe spanned exons 14 and 15 (HS01014856_m1). For any clinical screening test, it is suggested that more than one probe would need to be used. This was not possible during this project because of financial implications. A cut-off value of Rq = 0.6 was chosen to try and identify patients with clinically significant reduced gene transcription. It is important to be aware that this is an arbitrary value, and more accurate determination of a clinically significant threshold would need to be determined through large-scale studies.

The AI studies identified *APC* AI in 4 samples: 18% of the informative cohort. In 2 patients, Halo52 and Halo53, it appeared to be across the whole allele. In 2 patients, Halo08 and Halo51, it was definite at the final SNP only. When patients appeared to have lost a complete allele, this correlated with their gene expression results (Halo52 and Halo53). However, there was no correlation for those patients who appeared to have lost the 3' end of the allele. The *APC* Rq values for patients Halo08 and Halo51 were 0.869 and 0.826 respectively, so even if there is a reduction in transcription from one allele, the overall amount of cDNA present appears to be within the normal range. It is interesting to note that both Halo08 and Halo51 both have relatively strong phenotypes: Halo08 has 629 adenomas and Halo51 has 208 adenomas. In view of this, it might be expected that there could be reduced transcription from one allele, however the qPCR results did not support this finding, and the clinical implications of loss of the final SNP are uncertain.

3.4.2 Abnormal Splicing

cDNA analysis did not identify any splicing abnormalities in *APC* or *MUTYH.* However, the sample size was relatively small (45 patients for *APC* and 14 for *MUTYH)* and the patients' phenotypes were highly variable. It has previously been reported that 8% of NMI polyposis patients have an aberrant *APC* transcript (Spier *et al* 2012). However, all of Spier's index patients had at least 20 synchronous or 40 metanchronous adenomas. In the patient cohort for this study, it is not often clear from the patient notes whether the polyps were synchronous or metanchronous. The RNA samples obtained account for a maximum of 22 patients with at least 20 adenomas, so the phenotype of this study cohort is weaker.

cDNA sequencing for *APC* is a cheap technique to perform, so it is suggested that it forms part of the investigative protocol for NMI patients with an appropriate phenotype, for example those with at least 20 adenomas. *MUTYH* cDNA sequencing proved difficult to optimise. MAP is an AR disease, so even if any samples with abnormal *MUTYH* splicing had been identified, this alone would not be sufficient to explain the phenotype. Whilst *MUTYH* cDNA sequencing may be a useful tool in patients who are already known to carry one pathogenic *MUTYH* variant, it is unlikely to be cost and time efficient if it were performed on all NMI polyposis patients.

3.4.3 Ultradeep Sequencing

A total of 15 variants were identified for further investigation/ validation, 14 in *APC* and 1 in *MUTYH.* 4 had been previously validated by the NHS, including the monoallelic *MUTYH* mutation, which is known to be pathogenic.
Four of the *APC* variants were of very low frequency (present in <5% of reads), and one was at low frequency (present in 18% of reads). None of the 5 low frequency reads were validated. All of the variants which appeared to be approximately heterozygous were validated.

3.4.3.1 Validation of Low Frequency Variants

The Haloplex assay followed by UDS identified 5 very low/ low frequency variants. Whilst such variants may represent sequencing artefacts, they could also result

from the patient exhibiting generalised mosaicism. This study sought to validate such variants using COLD-PCR, which is a technique that can be used to enrich samples for low-abundance DNA mutations (Milbury *et al* 2011). The technique selectively amplifies low-abundance variants from mixtures of wild-type and mutant-containing sequences by using a critical denaturation temperature. A lower denaturation temperature allows the selective denaturation of amplicons which contain mutant/ wild type heteroduplexes or which contain Tm-reducing variants (Milbury *et al* 2011), so that they are preferentially amplified. It is reported that COLD-PCR followed by Sanger sequencing can be used to identify mutant fractions down to 0.1%, but it is noted that enrichment potential is highly dependent upon the type of mutation being analysed, the initial mutation abundance, and the PCR platform utilised (Milbury *et al* 2011).

None of the low frequency *APC* variants were definitely validated, although the result was ambiguous for sample Halo45. It is likely that they are sequencing artefacts from the UDS protocol, however it is also possible that the COLD-PCR reactions employed were not sufficiently sensitive to detect them. Fortunately, most of the variants examined were missense changes. Even if they had been real, such low frequency missense *APC* variants in an individual are unlikely to have a clinical impact (Cleary *et al* 2008). There was a stop-gain variant present in Halo63 which was not validated. However, it was only present in 2/11 reads and there was no evidence of it at all with Sanger sequencing, so it highly likely to be a sequencing artefact. If more sensitive validation of low frequency variants was required, it would be necessary to consider an alternative technique such as digital droplet PCR.

3.4.3.2 Validated Variants

A total of 10 potentially pathogenic *APC* or *MUTYH* variants were identified by UDS and validated with Sanger sequencing (Table 3.5) .

3.4.3.2.1 Validated Variants in *APC*

All exonic variants occurring in *APC* had previously been identified by the NHS diagnostic service. They were all putative missense variants which were of uncertain pathogenicity. Pathogenic *APC* mutations are usually truncating

mutations, and there is currently minimal evidence that missense variants are clinically significant (Cleary *et al* 2008).

Sample Halo23 has a variant in the untranslated region of exon 1. She has no evidence of AI, gene expression studies were normal with *APC* Rq = 0.804 and there was no evidence of abnormal splicing. Although the CADD score for this variant was 21.1, it has recently been classified as a benign variant on ClinVar (http://www.ncbi.nlm.nih.gov/clinvar/variation/133505/ accessed 3/5/16).

Halo35 has an *APC* intronic variant. Its CADD score is relatively low, at 18.02. There is no evidence of abnormal expression, with *APC* Rq = 0.894, and there was no evidence of abnormal splicing or allelic imbalance, so the variant is regarded as likely non-pathogenic.

Sample Halo46 has a 5'UTR variant and evidence suggestive of reduced *APC* transcription. She has been discussed above and this variant is believed to be pathogenic.

Sample Halo51 has an intronic *APC* variant, rs4705624. This has a CADD score of 16.36, and there is one submission on ClinVar (http://www.ncbi.nlm.nih.gov/clinvar/variation/82836/ accessed 01/03/16) in which it is given a classification of 'other' in the context of familial colorectal cancer. The patient exhibited loss of the final *APC* SNP, but there was no evidence of abnormal splicing. DNA was initially only available for the patient's mother: the variant was not present. However, once DNA was obtained for the patient's unaffected father, the same variant was present, so the variant has been regarded as non-pathogenic.

Sample Halo56 has an *APC* 5'UTR variant with a CADD score of 16.52. There is no evidence of reduced *APC* expression, with *APC* Rq = 0.793. AI studies are uninformative. Once the patient's parents had been recruited to the study the variant was identified in the unaffected father, so it has been regarded as non-pathogenic.

Sample Halo62 has an intronic variant with a CADD score of 18.98. There is no evidence of reduced *APC* expression, with *APC* Rq = 0.766, and AI studies show no evidence of AI. There is no abnormal splicing. Taking into account these findings, the variant has been assumed to be non-pathogenic.

3.4.3.2.2 Validated Variants in *MUTYH*

The final variant was identified in *MUTYH* in sample Halo40. This is one of the recurrent *MUTYH* mutations known to cause MAP, p.Gly396Asp. The patient is a carrier of a monoallelic mutation, so this by itself cannot explain the polyposis phenotype. The patient had no evidence of *MUTYH* AI, reduced gene expression or abnormal splicing.

3.4.3.3 Summary of Gene Capture and Ultradeep Sequencing

The Haloplex assay followed by UDS was used to look for variants throughout the whole of the *APC* and *MUTYH* genomic loci, and to identify any patients who might exhibit mosaicism. There are several other available techniques for targeted gene capture, including long-range PCR (LR-PCR). LR-PCR was initially trialled for use in this project. However, it failed to amplify the known mutant alleles from two FAP control patients. Despite carrying out a series of investigations to determine why this had occurred, the explanation was not clear. It is possible that certain mutations alter the structure of the DNA in such a way that the polymerase cannot replicate the mutant strand. Instead of LR-PCR, the Haloplex Assay was employed. The assay costs approximately £100 per patient sample, which is not prohibitive for use in a diagnostic setting. 60 patients underwent UDS of *APC* and *MUTYH.* All exonic variants which we filtered for had previously been identified by the NHS, so this study did not identify any mutations which had been 'missed' in the diagnostic setting. There were 6 variants outside of the ORF which were validated, one of which is believed to be pathogenic (the *APC* 5'UTR variant in Halo46). The remaining variants are regarded as unlikely to be clinically significant once the gene expression results, cDNA sequencing and AI studies are taken into consideration. This highlights the importance of using a range of investigative techniques, so that results are never viewed in isolation.

None of the patients had evidence of *APC* mosaicism. The depth of coverage was excellent with UDS, so if there were any true cases, it is unlikely to have been missed. It is reported that somatic mosaicism can occur in 10-20% of sporadic

cases of FAP (reviewed in Rohlin *et al* 2009; Hes *et al* 2008; Aretz *et al* 2007). 28 of the patients included in this study are thought to be sporadic polyposis cases, but the actual figure may differ from this as the family histories obtained are often vague/ missing.
It is also important to note that analysis has been performed in DNA extracted from whole blood. It is clear from the literature that some patients have mosaic mutations which are only present in colonic tissue (Jansen *et al* 2017), so we would not have identified these.

NGS inevitably generates large amounts of data, and a challenge is to identify what is clinically significant. In this thesis, CADD scoring was used as a tool to predict pathogenicity. CADD scoring gives a standardised genome-wide, variant scoring metric that incorporates the weighted results of widely used *in silico* pathogenicity prediction tools, such as SIFT and PolyPhen, and of genomic annotation sources like ENCODE (reviewed in van der Velde *et al* 2015). The resulting CADD scores are expressed as a measure of deleteriousness (selection pressure bias) for single-nucleotide variants (SNVs) and small indels (van der Velde *et al* 2015). Although CADD scores are useful in assisting the prioritisation of identified variants (van der Velde *et al* 2015), it is important to be aware that their sensitivity and specificity will never be 100%. CADD scores should be used in combination with other available data such as family segregation studies, functional analysis and literature reviews (Schiemann and Stowell 2016) and must not be regarded as the gold standard of variant appraisal.

3.4.4 Chapter Conclusions

60 NMI polyposis patients underwent interrogation of *APC* and *MUTYH* to try and determine the cause of their genetically unexplained phenotype. The study employed qPCR, AI screening, cDNA sequencing, karyotype analysis, MSP and UDS as research tools. Whilst the gene transcription and AI studies gave interesting results suggestive of reduced *APC* transcription for 4 patients, Halo46, Halo52, Halo53 and Halo64, the underlying cause for this was only determined for Halo46: UDS identified a 5'UTR variant in *APC*.

Although pathogenic variants/ molecular mechanisms of disease were not identified for the majority of the patients, the techniques used as part of this study are all

relatively cost-effective and could be considered for use in future diagnostic protocols for NMI polyposis patients, although further validation in clinical cohorts would be necessary.

In Chapter 4 of this thesis, all 60 patients who underwent UDS of *APC* and *MUTYH* will be subject to UDS of a further 15 genes possibly implicated in the pathogenesis of colorectal neoplasia to further investigate the cause of their phenotype. A subset of 24 patients will undergo whole exome sequencing (WES).

Chapter 4 Exome Sequencing in Polyposis Patients using Targeted and Whole Exome Approaches

4.1 Introduction

After a detailed examination of *APC* and *MUTYH,* the next phase of this study focused on exome sequencing. The initial strategy was to use a targeted approach. Targeted exome sequencing of 15 genes associated with colorectal/ intestinal neoplasia was performed. Subsequently, whole exome sequencing (WES) was undertaken for a subset of patients.

60 NMI patient samples and 11 relatives/ healthy controls underwent targeted exome sequencing of 15 genes: *AXIN2, BMPR1A, CDH1, CHEK2, EPCAM, MLH1, MSH2, MSH6, PMS2, POLD1, POLE, PTEN, SMAD4, STK11* and *TP53.* The relatives/ healthy controls were included to assess whether variants identified in NMI patients were occurring *de novo* and to allow variants to be assigned as likely non-pathogenic if they were found in a healthy control, once the mode of inheritance was taken into consideration. Gene capture was achieved using the Haloplex Assay and sequencing was performed using the HiSeq 2500 (Illumina). The patients were also screened for recurrent pathogenic *NTHL1* mutation, c.268A>T , p.Gln90* (Weren *et al* 2015). *NTHL1* had not been included in the Haloplex assay as the assay been designed prior to the publication of Weren's paper. Putatively pathogenic variants were validated with Sanger Sequencing and confirmed to be present in cDNA.

WES was performed on 24 patients. These were selected based on their phenotype, the availability of family members and the results of previous investigations completed as part of this study. Putatively pathogenic variants were validated with Sanger sequencing.

4.2 Targeted Exome Screening and Sequencing

Targeted exome sequencing used the Haloplex Assay for gene capture. The assay was designed in 2014, and was based on one of the commercially available CRC

gene panels available at the time, produced by AmbryGen: the Colonext NGS panel (http://www.ambrygen.com/tests/colonext accessed 20/4/16). 14 of the genes included on the study plate were present on the Colonext NGS plate (*BMPR1A, CDH1, CHEK2, EPCAM, MLH1, MSH2, MSH6, PMS2, POLD1, POLE, PTEN, SMAD4, STK11*, *TP53)*. *AXIN2* was added based on current evidence that it may have a role in the pathogenesis of colorectal neoplasia (Lammi *et al* 2004; Marvin *et al* 2011; Rivera *et al* 2014). The rationale for all of the genes investigated as part of this chapter is described below.

4.2.1.1 *NTHL1*

NTHL1 is mutated in *NTHL1*-associated polyposis/ CRC as described in Chapter 1 (1.5.1.4).

4.2.1.2 *BMPR1A* and *SMAD4*

Bone morphogenetic proteins (BMPs) are part of the TGFβ family. They bind to their receptors, triggering receptor activation, which proceeds to phosphorylate a receptor-associated SMAD protein. This then complexes with SMAD4, which translocates to the nucleus to regulate gene transcription (reviewed in Hardwick *et al* 2008) and plays a role in growth inhibition (reviewed in Fleming *et al* 2013). The *BMPR1A* gene encodes the BMP-receptor 1A, and *SMAD4* encodes the SMAD4 protein described above. Germline mutations in both genes are associated with impaired growth inhibition and are found in patients with Juvenile Polyposis Syndrome (JPS), in which patients develop intestinal hamartomas and have an increased risk of CRC (reviewed in Hardwick *et al* 2008). Germline *BMPR1A* mutations have also been identified in a 58-year old male patient with 'multiple' (8) colorectal adenomas (Lipton *et al* 2003b) and in families with microsatellite stable hereditary non-polyposis CRC (Nieminen *et al* 2011).

4.2.1.3 *CDH1*

The *CDH1* gene codes for E-cadherin, a member of the cadherin family of cell surface glycoproteins (reviewed in Kim *et al* 2000). It has roles in embryogenesis, polarisation, differentiation and cell migration in inflamed tissue (reviewed in Kim *et*

al 2000). Importantly, it is also known to be involved in intestinal cancer: loss of E-cadherin function is associated with invasiveness, lymph node metastasis and distant metastasis (reviewed in Kim *et al* 2000). It is well-established that germline mutations in *CDH1* result in an increased risk of developing hereditary diffuse gastric cancer (HDGC) and breast cancer (Pharoah *et al* 2001). It has also been suggested that inherited *CDH1* mutations might increase the risk of an individual developing CRC (Richards *et al* 1999; Kim *et al* 2000).

4.2.1.4 *CHEK2*

The protein product of the *CHEK2* gene acts as a checkpoint kinase (reviewed in Cybulski *et al* 2004). Activation of CHEK2 in response to DNA damage prevents the cell from undergoing mitosis (reviewed in Cybulski *et al* 2004). It is clear that *CHEK2* variants can play a role in the pathogenesis of breast cancer (Walsh *et al* 2006; Cybulski *et al* 2004), and there is evidence that *CHEK2* variants may also be associated with malignancies in multiple other organ systems, including the thyroid gland, prostate, colorectum and kidney (Cybulski *et al* 2004). In 2007, Cybulski *et al* (Cybulski *et al* 2007) reported that carriers of a *CHEK2* missense variant, p.Ile157Thr, had an increased risk of CRC compared to a control cohort (OR = 1.5, p = 0.002), and that this specific variant might be responsible for 3% of all CRC in Poland. Very similar results regarding increased CRC risk have also been reported in more recent papers investigating Russian populations (Yanus *et al* 2018), Polish populations (Suchy *et al* 2010) and in Finnish populations (Kilpivaara *et al* 2006).

4.2.1.5 *EPCAM* and the MMR Genes: *MLH1, MSH2, MSH6, PMS2*

The MMR system involves 7 key genes: *MSH2, MSH6, MSH3, MLH1, PMS1, PMS2* and *MLH3.* The role of five of these genes in polyposis, CRC and Lynch Syndrome has been described in Chapter 1 (1.3 and 1.5.1.5).

EPCAM is not a MMR gene, rather it is a cell adhesion molecule. However, it too plays an important role in the pathogenesis of Lynch Syndrome. Patients with germline deletions at the 3' end of the *EPCAM* gene experience epigenetic silencing of *MSH2,* which is 17kb downstream of *EPCAM* (reviewed in Tutlewska *et al* 2013).

4.2.1.6 *POLE* and *POLD1*

POLE and *POLD1* are mutated in PPAP and have been described in Chapter 1 (1.5.1.3).

4.2.1.7 *PTEN*

PTEN is a tumour suppressor gene: its product is a phosphatase which negatively regulates the phosphatidylinositol 3-kinase-AKT and mammalian target of rapamycin (mTOR) signalling pathways, which are involved in cell proliferation, cell cycle progression and apoptosis (reviewed in Lam-Himlin *et al* 2014). *PTEN* mutations are involved in the pathogenesis of several carcinomas, including breast, endometrial, thyroid, large bowel and kidney (reviewed in Lam-Himlin *et al* 2014). *PTEN* mutations are identified in 80% of patients with Cowden's Syndrome (reviewed in Omunsden and Lam 2012), described in Chapter 1 (1.5.3.3).

4.2.1.8 *STK11*

The *LKB1/ STK11* gene is a nuclear serine threonine kinase which regulates cell polarisation, growth and metabolism. Mutations in the gene account for approximately 50% of cases of PJS (reviewed in Omunsden and Lam 2012), described in Chapter 1 (1.5.3.1).

4.2.1.9 *TP53*

The p53 protein is a transcription factor which has a vital role in maintaining genomic stability (reviewed in Sarasqueta *et al* 2013). Following DNA damage, p53 activation causes arrest of the cell cycle to allow DNA repair (reviewed in Sarasqueta *et al* 2013). If the damage is too extensive, p53 can drive a cell towards senescence or apoptosis (reviewed in Sarasqueta *et al* 2013; reviewed in Vogelstein and Kinzler 2004). Somatic *p53* mutations are a key event in the development of CRC (reviewed in Iacopetta 2003; Vogelstein *et al* 1988). Germline *p53* mutations give rise to Li Fraumeni Syndrome, a highly penetrant cancer predisposition syndrome, in which patients develop a variety of early onset tumours, including sarcomas, carcinomas, haematological malignancies and brain tumours

(reviewed in Gonzalez *et al* 2009). CRC is observed in 2.8% of patients with germline *p53* variants (Gonzalez *et al* 2009).

4.2.1.10 *AXIN2*

It is well established that the majority of CRC are characterised by activation of the Wnt signalling pathway. As described in the previous chapters, nuclear β-catenin complexes with DNA-binding proteins of the TCF/ LEF family, which leads to enhanced expression of oncogenes involved in tumour progression (reviewed in Wu *et al* 2012). AXIN2 is a scaffold protein which is involved in regulating Wnt signalling. It forms part of the 'destruction complex' which targets β-catenin for degradation: AXIN2 supports the GSK3β-dependent phosphorylation of β-catenin, which marks the protein for ubiquitination and subsequent proteosomal degradation (reviewed in Wu *et al* 2012). *AXIN2* is a transcriptional target of β-catenin dependent Wnt signalling, and AXIN2 levels are elevated in cancers with Wnt activating mutations, therefore potentially negatively regulating Wnt signalling (reviewed in Mazzoni and Fearon 2014). There is some evidence that *AXIN2* mutations may have a role in colorectal neoplasia: this will be described in detail in Chapter 5.

4.2.2 Aims and Objectives: Targeted Exome Sequencing

The aim of the first part of this chapter was to identify novel genetic variants in genes thought to be/ known to be associated with intestinal neoplasia in NMI patients with multiple colorectal polyps. This involved screening for the known pathogenic mutation in *NTHL1* using Sanger sequencing and examining the exons of 15 genes using NGS technologies.

The polyposis patients included in this study are the same 60 patients that underwent UDS of *APC* and *MUTYH* described in Chapter 3, as are the 11 unaffected controls (3.2.5).

4.2.3 Methods

4.2.3.1 Screening *NTHL1*

DNA was extracted from whole blood of 60 NMI patients using standard protocols (2.3.2). PCR and Sanger sequencing was performed using standard reagents and reaction conditions (2.3.8.1, 2.3.8.5, 2.3.9-2.3.13) to screen for the recurrent polyposis-associated mutation in *NTHL1*, c.268A>T, p.Gln90*. The primers were designed by Dr. Laura Thomas (Appendix 4.1) who carried out the PCR reactions and sequencing.

4.2.3.2 Haloplex Assay and Sequencing

4.2.3.2.1 Targeted Exome Capture, Sequencing and Data Analysis

Targeted exome capture was achieved using the Haloplex assay (Agilent). The probes for the Haloplex assay were designed using Agilent's Sure Design software (https://earray.chem.agilent.com/suredesign/home.htm). DNA extracted from blood samples underwent UDS using the HiSeq 2500, as described in Chapter 2 (2.4.2). Bioinformatic analysis followed the protocol detailed in Chapter 3 (2.4.3).

The data was filtered to select for variants which were:

- rare, with an allele frequency </= 1% based on data from The 1000 Genomes Project
- not synonymous
- not known to be benign when investigated using 4 public databases (dbSNP, HGMD, InSiGHT and LOVD)
- present in all affected relatives
- not present in unaffected relatives (where family members were available for analysis)/ not present in healthy controls (the pattern of inheritance was taken into account. For example, when modelling dominant traits, variants which were present in a healthy control were excluded; when modelling autosomal recessive traits, variants identified in polyposis patients were only excluded if they were also homozygous in a healthy control)
- present in </= 3 samples

Variants were then analysed using CADD software (http://cadd.gs.washington.edu/home). Variants which had a CADD score >/= 15 were validated with Sanger sequencing provided they appeared real when observed using Integrative Genomics Viewer (IGV) (https://www.broadinstitute.org/igv/).

4.2.3.3 Validation of Identified Variants

Variants identified were validated with Sanger Sequencing, using standard reagents and reaction conditions (2.3.8.1, 2.3.8.5, 2.3.9-2.3.13). Variants present at a low frequency (<20%) were validated with COLD-PCR, using standard protocols (2.3.8.3, 2.3.8.4, 2.3.8.6, 2.3.8.7).

The primers were designed by Marc Naven (Cardiff University) using a combination of his own scripts, samtools and primer3 software, and by myself using primer3 software. They were supplied by Eurofins (Appendix 4.2).

4.2.3.4 Confirmation of the Variants in cDNA

RNA was converted to cDNA following standard protocols (2.3.3.2). DNA underwent PCR and Sanger sequencing as per the standard protocol (2.3.8.1, 2.3.8.5, 2.3.9-2.3.13). The primers used are listed in Appendix 4.3.

4.2.4 Results

4.2.4.1 Screening for the common mutations in *NTHL1*

60 patients were screened for the known pathogenic *NTHL1* mutation, c.268A>T, p.Gln90*. This was not identified in any of the patient samples.

4.2.4.2 Haloplex Assay and Sequencing

4.2.4.2.1 Metrics of UDS

The mean depth of coverage was 2900 reads across all 15 genes. The mean on-target coverage across all the genes was 98.1%. The details for each gene are listed in Table 4.1:

Gene	**Mean Depth of Coverage (number of reads)**	**Mean On-Target Coverage (%)**
AXIN2	2928	100
BMPR1A	3010	99.6
CDH1	3627	99.7
CHEK2	2232	89.3
EPCAM	2343	100
MLH1	2876	94.8
MSH2	2290	97.2
MHS6	3523	100
PMS2	2605	98.1
POLD1	3526	99.9
POLE	3910	99.8
PTEN	1834	97.6
SMAD4	2649	99.9
STK11	3112	99.8
TP53	3039	95.5

Table 4.1 Coverage of the 15 genes on the Haloplex assay

4.2.4.2.2 Variants Selected for Validation

60 NMI patients underwent targeted exome capture and UDS. A total of 33 different variants were selected for validation across 32 patients (Table 4.2). The validation sequencing traces are in Appendix 4.4.

Sample	Gene	Location	Variant	Frequency	CADD Score	Amino Acid Change	Genomic Annotation	Validated
Halo05	*CHEK2*	22: 29091740	C>G	1335/3023 (44%)	26.9	p.Arg406Pro	nonsynonymous SNV	Yes
Halo06	*POLD1*	19: 50918229	G>A	1716/3706 (46.4%)	22.4	p.Arg849His	nonsynonymous SNV	Yes
	MSH2	2:47639637	T>A	3/3 (100%)	24.5	p.Leu244Met	nonsynonymous SNV	No
Halo07	*SMAD4*	18:48603114	C>A	9/890 (1.01%)	24.7	p.Pro472Gln	nonsynonymous SNV	No
Halo14	*CHEK2*	22:29107974	C>T	277/632 (43.8%)	24.8	p.Glu239Lys	nonsynonymous SNV	Yes
Halo15	*POLE*	12:133249812	T>C	22813/45964 (49.6%)	24.4	p.Met471Val	nonsynonymous SNV	Yes
Halo17	*TP53*	17:7577117	A>T	55/3547 (1.56%)	29.8	p.Val274Asp	nonsynonymous SNV	No
Halo18	*POLE*	12:133202816	C>T	310/609 (50.9%)	23.8	p.Glu2140Lys	nonsynonymous SNV	Yes
Halo20	*MSH6*	2:48026861	C>T	7/669 (1.05%)	32	p.Ser580Leu	nonsynonymous SNV	No

Sample	Gene	Location	Variant	Frequency	CADD Score	Amino Acid Change	Genomic Annotation	Validated
Halo25	*CDH1*	16:68867388	G>A	1856/3910 (47.5%)	27.8	p.Gly879Ser	nonsynonymous SNV	Yes
Halo26	*MSH2*	2:47630458	A>G	439/858 (51.2%)	26.2	p.Tyr43Cys	nonsynonymous SNV	Yes
Halo27	*POLD1*	19:50919693	C>T	1081/2232 (48.5%)	34	p.Thr954Met	nonsynonymous SNV	Yes
Halo28	*AXIN2*	17:63534353	T>C	1147/3499 (32.8%)	15.42	p.Ser390Gly	nonsynonymous SNV	Yes
	MSH2	2:47639633	C>A	8/767 (1.04%)	25	p.Asn242Lys	nonsynonymous SNV	No
Halo29	*PTEN*	10:89720870	T>G	264/2436 (10.84%)	27.9	p.Phe341Val	nonsynonymous SNV	No
Halo31	*POLE*	12:133245032	A>T	2428/4862 (49.9%)	23	p.Phe695Ile	nonsynonymous SNV	Yes
Halo35	*CHEK2*	22:29121042	G>A	152/3145 (4.83%)	16.85	p.Thr215Ile	nonsynonymous SNV	No
Halo43	*PTEN*	10:89720678	A>T	12/848 (1.41%)	25.1	p.Thr277Ser	nonsynonymous SNV	No
Halo45	*MSH6*	2:48026228	C>T	580/1532 (37.9%)	29.5	p.Thr369Ile	nonsynonymous SNV	Yes

Sample	Gene	Location	Variant	Frequency	CADD Score	Amino Acid ChanGlye	Genomic Annotation	Validated
Halo46	*MLH1*	3:37067140	G>T	7/619 (1.13%)	39	p.Gly351*	stopgain	No
Halo47	*AXIN2*	17:63533512	C>A	357/742 (48.2%)	37	p.Gly548*	stopgain	Yes
Halo49	*MSH6*	2:48010592	G>T	8/734 (1%)	36	p.Gly74*	stopgain	No
Halo51	*PMS2*	7:6045634	T>C	936/2788 (34%)	25.6	p.Ile18Val	nonsynonymous SNV	No
Halo54	*CHEK2*	22:29090061	G>A	71/4620 (1.5%)	34	p.Arg474Cys	nonsynonymous SNV	Yes
Halo58	*MSH6*	2:48030603	C>T	44/4148 (1%)	23.8	p.Pro1073Ser	nonsynonymous SNV	No
Halo61	*POLE*	12:133245452	T>C	3617/6785 (53%)	26.0	p.Tyr623Cys	nonsynonymous SNV	Yes
Halo62	*CDH1*	16:68855966	G>A	609/1192 (51%)	23.9	p.Ala592Thr	nonsynonymous SNV	Yes
	EPCAM	2:47612347	G>A	267/922 (29%)	29.1	p.Glu301Lys	nonsynonymous SNV	Yes
Halo64	*BMPR1A*	10:88677029	T>G	19/1240 (1.5%)	29.7	p.Phe272Val	nonsynonymous SNV	No

Sample	Gene	Location	Variant	Frequency	CADD Score	Amino Acid ChanGlye	Genomic Annotation	Validated
Halo68	*AXIN2*	17:63533512	C>A	1176/2351 (50%)	37	p.Glylu548*	stopgain	Yes
Halo69	*POLE*	12:133253974	C>T	1611/3091 (52%)	24.4	p.Arg259His	nonsynonymous SNV	Yes
Halo70	*MSH6*	2:48030669	C>T	398/997 (40%)	35	p.Arg1095Cys	nonsynonymous SNV	Yes
Halo72	*TP53*	17:7578245	G>A	763/1656 (46%)	16.71	p.Arg202Cys	nonsynonymous SNV	Yes
Halo80	*TP53*	17:7578388	C>T	56/162 (35%)	28.5	p.Arg181His	nonsynonymous SNV	Yes
Halo81	*POLE*	12:133253974	C>T	1851/3784 (48.9%)	24.4	p.Arg259His	nonsynonymous SNV	Yes

Table 4.2 Variants selected for validation following targeted exome capture

4.2.4.2.3 Confirmation of the Variants in cDNA

At the time of the experiments, RNA was available for 14 patients with variants validated in DNA. All the variants which were identified in DNA were confirmed to be present in cDNA (Appendix 4.5).

4.2.5 Further Assessment of Validated Variants

Validated variants were checked in dbSNP (https://www.ncbi.nlm.nih.gov/projects/SNP/), HGMD (https://portal.biobase-international.com/hgmd/pro/start.php), InSiGHT (https://www.InSiGHT-group.org/variants/databases/) and LOVD (http://lovd.nl/3.0/home). International experts who specialise in specific genes were contacted to determine whether they had any further information regarding the clinical significance of the confirmed variants. I also liaised with AmbryGen, a company which provides a clinical genetics testing and diagnostic service in the United States of America.

Several of the variants identified using the Haloplex assay followed by UDS have been deemed to be pathogenic, but in clinical situations other than colorectal polyposis. For example, the *CHEK2* mutation found in Halo14 (c.715G>A:p.Gln239Lys) is assigned to be DM (pathological) in prostate cancer (CM030421: inherited variant) on HGMD and the *TP53* mutation in Halo072 (c.604C>T:p.Arg202Cys) is reported as DM in adrenocortical carcinoma (CM121764: inherited variant).

With regards to the *CDH1* variants, I contacted Professor Seruca's research group in Portugal. One of the *CDH1* variants which had been found in Halo62, c.1774G>A:p.Ala592Thr, has been classified as non-pathogenic. The group had already evaluated its significance *in vivo* and demonstrated that cells expressing the variant are not invasive (Keller *et al* 2004). The second variant identified in Halo25, c.2635G>A:p.Gly879Ser, had been reported in an 81-year old Caucasian woman with lobular breast cancer, but its functional relevance remains to be evaluated. Professor Seruca noted that although CRC can occur as part of the tumour spectrum in Hereditary Diffuse Gastric Cancer, which can develop as a consequence of germline mutations in *CDH1,* it is specifically of signet ring cell type. She reported that there is no evidence to suggest that *CDH1* mutations increase the risk of CRC associated with colorectal polyposis (personal communication

10/10/2016). In light of this, the *CDH1* mutations were not pursued as part of this project.

AmbryGen provided their classification, if available, of the validated variants (VLB: variant likely benign, Poly: polymorphism, VUS: variant of unknown significance, VLP: variant likely pathogenic). This is listed in Table 4.3:

Gene	Transcript Reference Sequence	Location	Variant	Amino Acid Change	Genomic Annotation	Classification
CHEK2	NM_007194	22: 29091740	C>G	p.Arg406Pro	nonsynonymous SNV	Not reported
CHEK2	NM_007194	22:29107974	C>T	p.Glu239Lys	nonsynonymous SNV	VUS
CHEK2	NM_007194	22:29090061	G>A	p.Arg474Cys	nonsynonymous SNV	VUS
POLD1	NM_002691	19: 50918229	G>A	p.Arg849His	nonsynonymous SNV	Poly
POLD1	NM_002691	19:50919693	C>T	p.Thr954Met	nonsynonymous SNV	Not reported
POLE	NM_002691	12:133245032	A>T	p.Phe695Ile	nonsynonymous SNV	Poly
POLE	NM_002691	12:133245452	T>C	p.Tyr623ys	nonsynonymous SNV	VUS
POLE	NM_006231	12:133249812	T>C	p.Met471Val	nonsynonymous SNV	Not reported
POLE	NM_006231	12:133202816	C>T	p.Gln2140Lys	nonsynonymous SNV	VUS
POLE	NM_006231	12:133253974	C>T	p.Arg259His	nonsynonymous SNV	VLB

CDH1	NM_004360	16:68867388	G>A	p.Gly879Ser	nonsynonymous SNV	VLB
CDH1	NM_004360	16:68855966	G>A	p.Ala592Thr	nonsynonymous SNV	Poly
MSH2	NM_000251	2:47630458	A>G	p.Tyr43Cys	nonsynonymous SNV	VUS
MSH6	NM_000179	2:48026228	C>T	p.Thr369Ile	nonsynonymous SNV	VUS
MSH6	NM_000179	2:48030669	C>T	p.Arg1095Cys	nonsynonymous SNV	VUS
AXIN2	NM_004655	17:63534353	T>C	p.Ser390Gly	nonsynonymous SNV	Not reported
AXIN2	NM_004655	17:63533512	C>A	p.Glu548*	stopgain	Not reported
EPCAM	NM_002354	2:47612347	G>A	p.Gln30Lys	nonsynonymous SNV	Not reported
TP53	NM_000546	17:7578245	G>A	p.Arg202Cyus	nonsynonymous SNV	VUS
TP53	NM_000546	17:7578388	C>T	p.Arg181His	nonsynonymous SNV	VLP

Table 4.3: AmbryGen classification of validated variants

Based on this information, the following conclusions were drawn:

4.2.5.1 *CHEK2*

It is possible that *CHEK2* variants may have a role in colorectal neoplasia. Three different variants were validated in this cohort of NMI patients: c.1217G>C:p.Arg406Pro in Halo05, c.715G>A:p.Glu239Lys in Halo14 and c.1420C>T:p.Arg474Cys in Halo54. However, during the time course of this project it was not possible to recruit additional family members for *de novo*l segregation analysis. Therefore these variants have not been pursued in the short-term.

4.2.5.2 *POLE* and *POLD1*

The VUS/ unreported *POLE* variants were selected for additional studies in Chapter 5 of this thesis. One of the *POLD1* variants, c.2546G>A:p.Arg849His, is recorded as a polymorphism by AmbryGen, and has recently been classified as benign in dbSNP (https://www.ncbi.nlm.nih.gov/clinvar/variation/220865/ accessed 08/05/2018) so was not pursued. The second *POLD1* variant, c.2861C>T:p.Thr954Met, is not previously reported. It was taken forward for further work described in Chapter 5.

4.2.5.3 *CDH1*

These variants were not investigated following Professor Seruca's advice that they are unlikely to be pathogenic in the context of polyposis. This was also consistent with AmbryGen's classification of the variants identified as VLB/ Poly.

4.2.5.4 *MSH2, MSH6, MSH6, MLH1, EPCAM*

Variants in the MMR genes were selected for further work. Although the AmbryGen classification was taken into consideration, the gold standard classification is that provided by the InSiGHT consortium (https://www.InSiGHT-group.org/syndromes/lynch-syndrome/), so this was given the greatest weight. The only variant which wasn't pursued was the *EPCAM* missense variant. Only deletions of *EPCAM* that inactivate the adjacent *MSH2* gene in Lynch syndrome are considered pathogenic, (Tutlewska *et al* 2013) rather than missense changes.

4.2.5.5 *AXIN2*

Neither of the variants identified had been reported by AmbryGen. We were unable to recruit further relatives for Halo28, and the variant was a missense change, so is less likely to be pathogenic. Therefore, the c.1168A>G:p.Ser390Gly variant was not taken forward. The variant found in Halo47 and Halo68, c.1642G>T:p.Glu548*, was extensively investigated as reported in Chapter 5 of this thesis.

4.2.5.6 *TP53*

One of the variants has been discussed above. *TP53* c.604C>T:p.Arg202Cys was identified in Halo72, and is thought to have a role in adrenocortical carcinoma. Halo72 had an interesting family history, including several family members in the preceding generation having CRC/ lung cancer/ cancer of unknown origin. However, we were unable to recruit any further family members, so the variant has not been pursued at the present time. The second variant, c.542G>A:p.Arg181His, is recorded as likely pathogenic by AmbryGen, who have 12 cases of patients carrying this variant. 11/12 had breast cancer, and the remaining patient had a family history of breast cancer. Although most of the families had histories of further malignancies, they did not fulfil Li Fraumeni diagnostic criteria, and there were no cases of polyposis. In light of this, the variant was not investigated further.

The variants which were selected for further investigations were therefore those whose significance was unknown/ which were unreported and for which basic functional/ *in silico* work could be carried out and/ or those which were identified in patients in whom relatives were available to be recruited for genetic analysis to determine *de novo* status or segregation.

The genes I selected to focus on were the MMR genes, *POLE, POLD1* and *AXIN2*. This work is described in Chapter 5 of this thesis.

4.3 Whole Exome Sequencing

The next phase of this project involved whole exome sequencing (WES) of 24 patients, plus 5 relatives. The relatives were included to assess whether variants identified in NMI patients were occurring *de novo* and to allow variants to be assigned as likely non-pathogenic if they were found in a healthy control, once the mode of inheritance was taken into consideration. Although most of the patients included in this thesis will ultimately undergo WES and/ or whole genome sequencing, this cohort was selected based on a phenotype of early onset or profuse polyposis and/ or the availability of relatives and/ or those patients in whom a novel putative polyposis-causing variant had already been identified, to help exclude the presence of an additional previously undetected polyposis-associated variant.

The patients and relatives included are recorded in Table 4.4 below:

Unique Identifier	**Demographic Details**	**Clinical Phenotype**	**Relative Available**
Halo05	46-year-old female	Aged 39 had multiple adenomas (at least 19): one polyp contained Duke's pT1 CRC	
Halo08	70-year-old female	629 adenomas	
Halo15	84-year-old male	12 adenomas	
Halo17	66-year-old male	25 mixed polyps: adenomas and hyperplastic lesions (exact numbers of each is not clear). 2 CRC aged 63 (pT2,pT3)	

Halo18	72-year-old male	Referred from screening programme. 27 polyps: adenomas with 4 HPPs	
Halo19	62-year-old female	>10 adenomas and CRC (no further information available about the CRC)	
Halo27	67-year-old female	17 adenomas	
Halo28	61-year-old male	1990: sigmoid colectomy for a probable diverticular perforation. Multiple 'benign metaplastic polyps' were noted ?number. 2012: subtotal colectomy for >30 polyps. The majority of the polyps were TA; 2 were serrated adenomas with LG dysplasia.	
Halo40	70-year-old female	18 adenomas, 3 HPPs	
Halo45	77-year-old man	27 adenomas, 7 HPP	
Halo46	44-year-old female	Clinical FAP – has had a colectomy and proctectomy	
Halo47	51-year-old female	Colonoscopy found in excess of 50 sessile polyps throughout the colon. 2 of these were biopsied: TVA. Px had	Halo68 (Sister) 3 TVA with LGD, 10 TA with LGD and 7 HPPs.

		a subtotal colectomy - ?4 polyps were sampled: 2 TA, 2 'serrated adenomas'. The patient subsequently had a completion colectomy – she was found to have a rectal carcinoma (pT4N2Mx) Sister of Halo68	
Halo48	71-year-old male	Referred from screening. 30 polyps: 4 have been biopsied: all adenoma. Subsequent colectomy.	
Halo49	57-year-old female	Rectal adenocarcinoma age 32 (?stage), ameloblastoma age 38. 22 polyps: mixture of adenomas and hyperplastic lesions	
Halo51	43-year-old female	208 adenomas	Halo77 (mother): unaffected Halo78 (father): unaffected
Halo52	59-year-old male	Over 1000 adenomas	
Halo53	51-year-old male	Approximately 400 adenomas, carcinoma age 27 (Duke's C)	Halo75 (mother): unaffected Halo76 (father): unaffected
Halo55	54-year-old male	18 polyps: 9 adenomas, 9 serrated lesions	F_Halo55 (father): unaffected

Halo61	60-year-old female	60-year-old female with 11 polyps: 2 serrated adenomas, 7 TA, 1 TVA, and 1 HPP	
Halo63	68-year-old male	At least 24 polyps. 5 biopsies show 4 adenomas and 1 HPP	
Halo64	54-year-old female	Thousands of colorectal polyps, those which have been biopsied showed adenomas. Caecal carcinoma aged 23 (?stage)	
Halo65	33-year-old male	117 adenomas	
Halo66	70-year-old male	23 adenomas, 4 HPPs	

Table 4.4: Patients and relatives undergoing WES

4.3.1 Methods

DNA was extracted from whole blood by the AWMGS (2.3.2). The concentration of the DNA samples to be examined were determined using the Invitrogen Qubit kit following the manufacturer's protocol. The DNA was diluted to a concentration of 5ng/µl.

4.3.1.1 Sequencing

Sequencing was carried out by Sarah Edkins, Wales Gene Park, using the Nextera Rapid Capture Enrichment Library Preparation and Sequencing Protocols (2.4.4).

4.3.1.2 Sequence Analysis

Bioinformatic analysis was performed by Dr. Kevin Ashelford (Cardiff University) (2.4.5).

4.3.1.3 Variant Analysis and Selection

A candidate gene approach was initially adopted, and the data was filtered for genes which were:

1) present in an in-house list of 1177 candidate CRC genes (Smith *et al* Appendix 4.6) or

2) were established cancer predisposition genes (Rahman 2014) or

3) were genes which are differentially expressed in solid tumours (Digital Differential Display (DDD) genes, Scheurle *et al* 2000)

Novel genes outside of this list were being investigated by Beth Bradford as her professional training year (PTY) project (Bradford, B (2017) *Use of Whole Exome Sequencing for the Identification of Novel Genetic Mechanisms in Colorectal Polyposis*). Any patient not investigated by Bradford was subject to interrogation by myself, to ensure that all patient samples had been examined in the same way. I therefore examined samples Halo05, Halo15, Halo27, Halo28, Halo47, Halo61 and Halo68.

The complete protocol for shortlisting variants of interest is outlined in Figure 4.1.

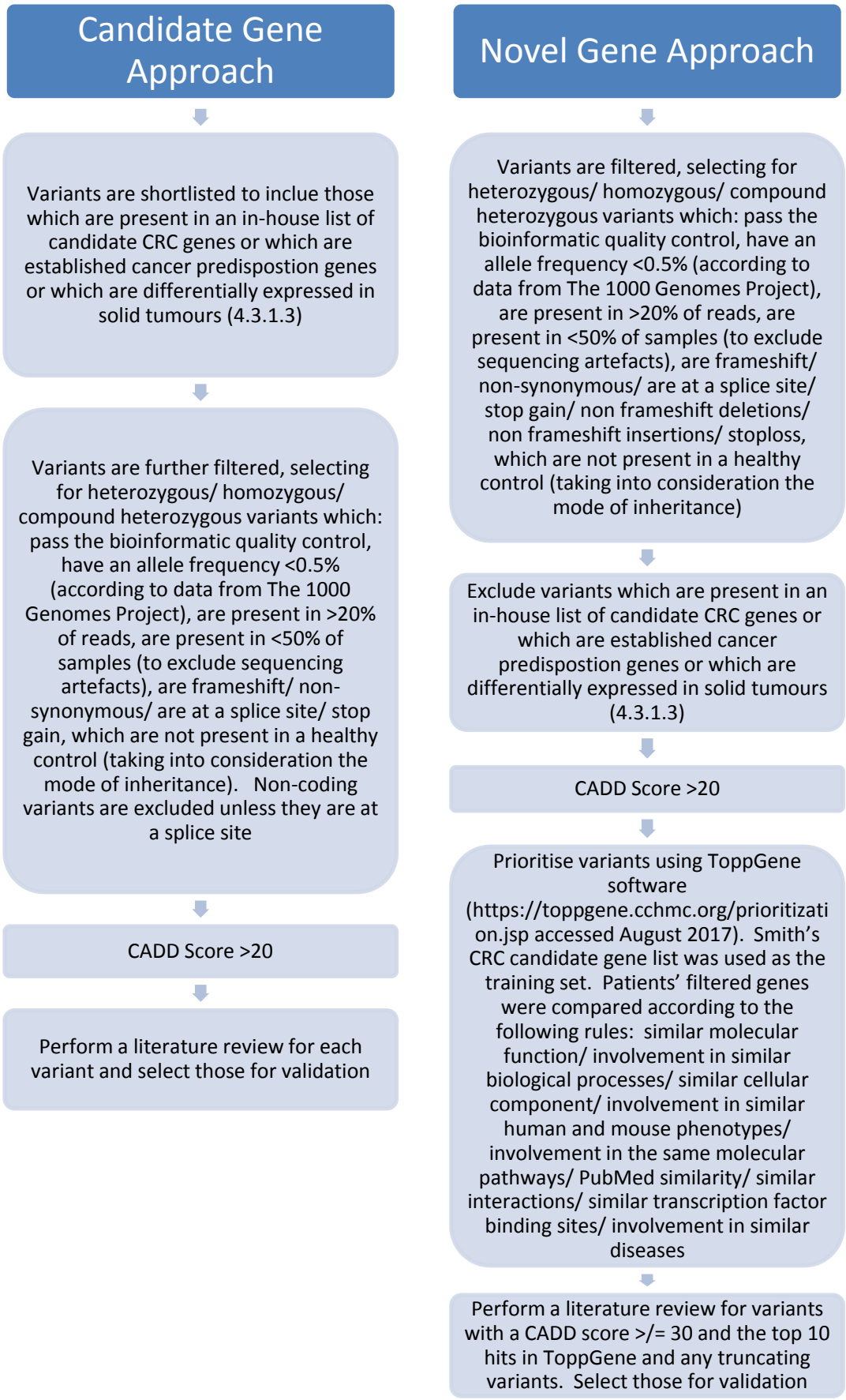


Figure 4.1 WES variant analysis and selection

4.3.1.4 Validation of Identified Variants

Variants were validated with Sanger sequencing, using standard reagents and reaction conditions (2.3.8.1, 2.3.8.5, 2.3.9-2.3.13). Primers were designed by Dr. Marc Naven (Cardiff University) using a combination of his own scripts, samtools and primer3 software, and by myself using primer3 software. They are listed in Appendix 4.7.

4.3.2 Results

4.3.2.1 Coverage

The mean depth of coverage across the 29 patients/ relatives was 61.99 reads. The details for each patient, including the percentage of target region covered at different depths, are recorded in Table 4.5:

		Percentage of target region covered at the following coverage depths:						
Sample	**Mean Reads**	**1x**	**4x**	**10x**	**20x**	**30x**	**40x**	**50x**
Halo05	88.82	97.20	95.24	92.51	87.21	81.18	74.76	68.09
Halo08	50.36	97.26	92.08	81.69	67.66	56.34	46.60	38.23
Halo15	123.42	98.47	95.99	91.60	84.90	79.14	74.13	69.61
Halo17	49.62	97.88	93.33	83.02	67.78	55.27	44.74	36.01
Halo18	49.11	97.95	93.48	83.17	69.90	55.46	44.85	36.02
Halo19	44.73	97.36	91.86	80.32	64.60	52.03	41.55	33.01
Halo27	85.95	97.97	95.01	89.39	80.82	73.54	67.13	61.17
Halo28	64.59	97.26	94.97	91.20	83.39	74.30	64.67	55.17
Halo40	60.54	98.01	97.73	86.87	73.69	62.23	52.22	43.53

Halo45	40.69	97.42	91.46	79.15	64.40	49.11	38.20	29.52
Halo46	76.93	98.24	95.63	89.59	79.02	69.69	61.43	53.92
Halo47	73.33	97.93	95.68	92.77	86.91	79.50	71.04	62.15
Halo61	139.76	98.66	96.73	93.43	87.76	82.43	77.62	73.23
Halo68	74.90	97.25	95.11	91.90	85.40	77.84	69.70	61.41
Halo48	36.24	98.07	93.33	79.35	57.19	41.45	30.65	22.99
Halo49	64.64	98.26	95.19	87.78	75.11	64.14	54.52	46.03
Halo51	52.73	98.06	94.47	85.40	70.66	58.30	47.70	38.79
Halo77	51.57	97.70	93.77	84.16	69.62	57.62	47.27	38.49
Halo78	61.28	98.04	93.93	84.21	69.45	57.28	46.81	37.98
Halo52	62.20	98.26	95.27	87.78	74.86	63.69	53.94	45.40
Halo53	62.15	98.22	95.10	87.52	74.69	63.58	53.85	45.30
Halo75	38.16	96.85	90.30	77.47	60.55	46.94	35.84	27.21
Halo76	62.76	97.56	92.81	83.67	71.53	61.88	53.57	46.27
Halo55	45.34	97.56	92.19	80.85	65.29	52.89	42.50	33.87
Halo55 Father	56.64	98.90	94.08	85.05	71.50	60.30	50.63	42.21
Halo63	54.32	98.14	94.38	85.26	70.92	59.00	48.77	40.03
Halo64	58.35	98.05	94.42	85.95	72.61	61.50	51.83	43.31
Halo65	54.59	97.95	94.01	84.97	71.07	59.38	49.24	40.47
Halo66	14.02	93.60	76.76	50.51	24.75	12.15	5.83	2.71

Table 4.5: Patient-specific depth of coverage for WES

4.3.2.2 Variants Selected for Validation

106 variants were selected for validation across the 24 patients. The details and results of validation are in Table 4.6. The validation sequencing traces are in Appendix 4.8.

Patient ID	Gene	Chr.	Location	Variant	CADD Score	Validated	Genomic Annotation	Notes
Halo05	*LZTS1*	8	20107358	G>A	34	Yes	Missense	
Halo05	*ERCC6*	10	50690821	G>A	34	No	Missense	There was a synonymous variant adjacent to the called variant
Halo05	*LRP4*	11	46918516	G>A	29.3	Yes	Missense	
Halo05	*ATM*	11	108114679	G>T	24.4	No	Splice site acceptor	Only present in 2/5 reads (NGS)
Halo05	*ATM*	11	108114684	G>T	23.3	No	Missense	Only present in 2/7 reads (NGS)
Halo05	*ATM*	11	108114689	C>T	23.4	No	Missense	Only present in 2/10 reads (NGS)
Halo05	*ATM*	11	108196797	G>A	34	Yes	Missense	
Halo05	*PALB2*	16	23635370	C>T	26.6	Yes	Missense	
Halo05	*LIG1*	19	48620943	C>A	29	Yes	Missense	

Patient ID	Gene	Chr.	Location	Variant	CADD Score	Validated	Genomic Annotation	Notes
Halo05	*FZD5*	2	208633009	C>T	22.5	Yes	Missense	
Halo05	*ATR*	3	142272170	A>G	20.5	Yes	Missense	
Halo05	*MCPH1*	8	6479113	C>T	24.6	Yes	Missense	
Halo08	*EYA4*	6	133789765	C>T	24.8	Yes	Missense	
Halo08	*BCLAF1*	6	136599544	G>A	36	Yes	Stopgain	
Halo08	*BMP1*	8	22049596	G>A	33	Yes	Missense	
Halo15	*SFN*	1	27190196	A>T	26.1	Yes	Missense	
Halo15	*CELSR2*	1	109812092	G>A	22	Yes	Missense	
Halo15	*CAPN9*	1	230907799	C>T	35	Yes	Missense	
Halo15	*ATR*	3	142272170	A>G	20.5	Yes	Missense	

Patient ID	Gene	Chr.	Location	Variant	CADD Score	Validated	Genomic Annotation	Notes
Halo15	*ZKSCAN 4*	6	28219686	CGGTC A>C	24.4	No	Frameshift	There was low coverage at this locus. The variant was called in 2/3 reads (NGS)
Halo15	*DAAM2*	6	39864686	C>T	34	Yes	Missense	
Halo15	*EGFR*	7	55273086	G>A	35	Yes	Missense	
Halo15	*ST18*	8	53030923	G>T	31	Yes	Missense	
Halo15	*NCOR2*	12	124824869	G>A	24.3	Yes	Missense	
Halo15	*NCOR2*	12	124911260	C>T	24.6	Yes	Missense	
Halo15	*HDAC5*	17	42171169	G>A	25.6	Yes	Missense	
Halo17	*BMPR2*	2	203407059	G>A	23.1	Yes	Missense	
Halo17	*BMPER*	7	34125622	C>T	34	Yes	Missense	

Patient ID	Gene	Chr.	Location	Variant	CADD Score	Validated	Genomic Annotation	Notes
Halo17	*TTI2*	8	33361016	C>T	24.6	Yes	Missense	
Halo17	*CBL*	11	119169085	G>A	21.6	Yes	Missense	
Halo17	*DSC3*	18	28605748	C>T	35	Yes	Missense	
Halo17	*DSC2*	18	28647999	T>TTC	35	Yes	Frameshift	
Halo18	*CNKSR1*	1	26515380	G>A	33	Yes	Missense	
Halo18	*POLQ*	3	121207509	CAATAGTA>C	34	Yes	Frameshift	
Halo18	*CBL*	11	119148958	T>C	23.9	Yes	Missense	
Halo18	*RAD51B*	14	68353893	A>G	27.2	Yes	Missense	
Halo19	*THRAP3*	1	36757052	G>A	24.3	Yes	Missense	

Patient ID	Gene	Chr.	Location	Variant	CADD Score	Validated	Genomic Annotation	Notes
Halo19	*RB1*	13	49047524	G>A	28	Yes	Missense	
Halo19	*MGA*	15	42028820	A>G	22.2	Yes	Missense	
Halo19	*WNT9B*	17	44953675	G>A	31	Yes	Missense	
Halo19	*MED12*	X	70342603	G>A	33	Yes	Missense	
Halo27	*PLK3*	1	45267346	G>T	33	Yes	Missense	
Halo27	*RBM5*	3	50155887	TGA>T	35	No	Frameshift	There was relatively low coverage at this locus and the variant was only called in 5/22 reads (NGS)
Halo27	*TRRAP*	7	98591187	G>C	21.6	Yes	Missense	
Halo27	*KAT5*	11	65482096	G>A	28.1	Yes	Missense	
Halo27	*PTPRH*	19	55697712	G>A	36	Yes	Stopgain	

Patient ID	Gene	Chr.	Location	Variant	CADD Score	Validated	Genomic Annotation	Notes
Halo28	*PIK3C2A*	11	17140810	A>G	23.5	Yes	Missense	
Halo28	*CD82*	11	44626709	G>A	33	Yes	Missense	
Halo28	*RBMS2*	12	56982155	C>T	21.3	Yes	Missense	
Halo28	*PPP1R13B*	14	104205127	T>C	23.1	Yes	Missense	
Halo28	*MGA*	15	42058553	G>C	22	Yes	Missense	
Halo28	*DCC*	18	50961517	G>A	28.2	Yes	Missense	
Halo28	*BIRC6*	2	32706513	G>T	32	Yes	Missense	
Halo28	*EYA4*	6	133789765	C>T	24.8	Yes	Missense	
Halo28	*AKAP9*	7	91623985	G>C	26	Yes	Missense	

Patient ID	Gene	Chr.	Location	Variant	CADD Score	Validated	Genomic Annotation	Notes
Halo28	*MET*	7	116381047	A>G	23.1	Yes	Missense	
Halo40	*GDF7*	2	20870532	C>A	23.6	Yes	Missense	
Halo40	*PDCD6IP*	3	33883492	G>A	30	Yes	Missense	
Halo40	*LRP5*	11	68183958	G>A	23.9	Yes	Missense	
Halo40	*DSC2*	18	28672114	C>T	21.7	Yes	Missense	
Halo45	*EIF3I*	1	32688188	A>G	28.2	Yes	Missense	
Halo45	*HELQ*	4	84374567	C>T	25.5	Yes	Missense	
Halo45	*RPLP2*	11	810305	A>G	22.6	Yes	Missense	
Halo46	*MSH4*	1	76345823	A>G	28.6	Yes	Missense	

Patient ID	Gene	Chr.	Location	Variant	CADD Score	Validated	Genomic Annotation	Notes
Halo47	*PPP1R13B*	14	104245134	C>T	23.5	Yes	Missense	
Halo47	*XAB2*	19	7694391	G>C	23.5	Yes	Missense	
Halo47	*ENG*	9	130588091	C>T	23.9	Yes	Missense	
Halo48	*ADAM17*	2	9633092	C>T	34	Yes	Missense	
Halo48	*FZD3*	8	28420428	G>A	23	Yes	Missense	
Halo49	*BMPR1B*	4	96070060	G>C	33	Yes	Missense	
Halo49	*MSH3*	5	80063899	G>C	23.7	Yes	Missense	
Halo49	*EPHB4*	7	100421340	C>T	25.2	Yes	Missense	
Halo49	*ID1*	20	30193351	C>T	29	Yes	Missense	

Patient ID	Gene	Chr.	Location	Variant	CADD Score	Validated	Genomic Annotation	Notes
Halo52	*EP400*	12	132471141	C>T	15.34	Yes	Missense	
Halo52	*TSC22D1*	13	45008887	G>T	24.6	Yes	Missense	
Halo52	*FOXC2*	16	86602293	G>C	28.3	Yes	Missense	
Halo52	*PHF12*	17	27239701	T>G	18.87	Yes	Missense	
Halo55	*DVL1*	1	1273404	G>A	22.8	Yes	Missense	
Halo55	*BMP8B*	1	40228846	G>T	39	Yes	Stopgain	
Halo55	*TLE1*	9	84228372	G>A	29.5	Yes	Missense	
Halo61	*PARP1*	1	226564855	G>A	26.2	Yes	Missense	
Halo61	*ANAPC2*	9	140069828	C>T	31	Yes	Missense	

Patient ID	Gene	Chr.	Location	Variant	CADD Score	Validated	Genomic Annotation	Notes
Halo61	*RECQL*	12	21623219	G>C	38	Yes	Stopgain	
Halo61	*INHBE*	12	57850383	G>A	33	Yes	Missense	
Halo61	*POSTN*	13	38156538	C>T	34	Yes	Missense	
Halo61	*EEF2K*	16	22269048	C>T	26.8	Yes	Missense	
Halo61	*ZFP14*	19	36831616	T>A	25.9	Yes	Missense	
Halo63	*BIRC6*	2	32726929	A>C	22.3	Yes	Missense	
Halo63	*XRCC5*	2	217026733	G>A	24.1	Yes	Missense	
Halo63	*WNT10A*	2	219754822	G>A	24	Yes	Missense	
Halo63	*CHD7*	8	61734439	G>A	25.7	Yes	Missense	
Halo63	*WISP1*	8	134232908	C>T	24.3	Yes	Missense	

Patient ID	Gene	Chr.	Location	Variant	CADD Score	Validated	Genomic Annotation	Notes
Halo63	*ARID2*	12	46233172	C>T	26.4	Yes	Missense	
Halo64	*MTHFR*	1	11854085	T>A	23.3	Yes	Missense	
Halo64	*WAPAL*	10	88259879	T>C	25	Yes	Missense	
Halo65	*POLQ*	3	121207520	G>A	24.2	Yes	Missense	
Halo65	*ERCC8*	5	60194107	G>T	24.4	Yes	Missense	
Halo65	*NUDT1*	7	2284301	G>A	23.7	Yes	Missense	
Halo65	*ATM*	11	108186610	G>A	31	Yes	Missense	
Halo65	*TMBIM6*	12	50146761	C>T	24.7	Yes	Missense	
Halo66	*MSH3*	5	80063896	C>T	22.8	Yes	Missense	
Halo66	*MMS19*	10	99218456	C>T	32	Yes	Missense	

Patient ID	Gene	Chr.	Location	Variant	CADD Score	Validated	Genomic Annotation	Notes
Halo66	*XAB2*	19	7694391	G>C	23.5	Yes	Missense	
Halo68	*PPP1R13B*	14	104245134	C>T	23.5	Yes	Missense	
Halo68	*XAB2*	19	7694391	G>C	23.5	Yes	Missense	
Halo68	*ENG*	9	130588091	C>T	23.9	Yes	Missense	

Table 4.6: WES validation

All of the validated variants were subsequently investigated using multiple databases to determine whether there was any additional information available about their likely pathogenicity: dbSNP, Exac, Ensembl, HGMD, LOVD, COSMIC, CBioPortal, ClinVar, CanVar (Databases accessed May/ June 2017, http://exac.broadinstitute.org/, http://grch37.ensembl.org/index.html, https://portal.biobase-international.com/hgmd/pro/start.php, http://www.lovd.nl/3.0/home, http://grch37-cancer.sanger.ac.uk/cosmic, http://www.cbioportal.org/index.do, https://www.ncbi.nlm.nih.gov/clinvar/, https://canvar.icr.ac.uk/, https://genome.ucsc.edu/cgi-bin/hgTracks?db=hg19&lastVirtModeType=default&lastVirtModeExtraState=&virtModeType=default&virtMode=0&nonVirtPosition=&position=chr11%3A108186595-108186625&hgsid=594778521_y7cxR9MaVPipen9AAAExDrahzrKP).
The results are summarised in Appendix 4.9. Particular attention was paid to the clinical situations in which specific inherited variants had been identified in, and whether or not the variants had been described as a somatic change in different cancer types.

Relatives were available for Halo51, Halo53 and Halo55. For both Halo51 and Halo53, no putative pathogenic *de novo* variants were identified. For Halo55 only 2 were found: missense changes in *TLE1* and *BRIC6*. A third variant identified in Halo55 was a stopgain in *BMP8B.* At a relatively late stage in the project, DNA became available for the patient's unaffected mother, in whom the *BMP8B* variant was also identified, so it is unlikely to be clinically significant.

4.4 Discussion

4.4.1 Targeted Exome Sequencing

The Haloplex assay was used for targeted exome capture of 15 candidate polyposis genes, followed by UDS using the HiSeq 2500 (Illumina). 33 different variants across 32 patients were selected for validation (a total of 35 variants as 2 different variants were identified in 2 sets of different patients). 21 variants which appeared to be heterozygous changes were validated. A possible *MHS2* homozygous variant observed in patient Halo06 was not validated. There were only 3 reads with UDS, so this is highly likely to be an artefact. A *PMS2* variant identified in Halo51 in 34% of reads was not validated. There are known to be multiple pseudogenes of *PMS2* (Vaughn *et al* 2010), so it is likely that UDS had identified a variant in such a pseudogene.

Twelve low frequency variants underwent COLD-PCR protocols followed by Sanger sequencing. In 2 samples, Halo17 and Halo49, the COLD-PCR protocols used did not successfully amplify the desired amplicon, and the required melting temperature was 95^0C as in standard PCR. For sample Halo64 the full COLD PCR protocol appeared to work at denaturation temperatures as low as 68°C. However, this would be extremely unlikely, if not impossible. It is most probable that the hybridisation step did not work, therefore the effective denaturation temperature would have been 95°C.

11/12 low frequency variants were not validated, so the UDS results are likely to be artefacts. However, it is possible that the COLD-PCR was not sufficiently sensitive to detect them. COLD-PCR has been discussed in Chapter 3 (3.4.3.1). The one low frequency variant which was validated was identified in sample Halo54: *CHEK2* 22:29090061 G>A. The UDS results indicated a variant fraction of 1.5% (71/4620 reads). However, the trace obtained with the Sanger sequencing validation revealed that the patient was approximately heterozygous rather than a very-low frequency mosaic (Appendix 4.4), suggesting that there had been preferential amplification of the wild-type allele with UDS.

The aim of this part of the project was to identify novel/ rare variants in genes already known to have/ potentially have a role in colorectal neoplasia. A key issue encountered was the lack of availability of relatives for many of the probands for determination of the variant's *de novo* status or assessment of whether it was segregating with the disease. Without such evidence, it is often not sensible to commence functional studies. The variants that have been selected for further work in this thesis are a novel truncating *AXIN2* mutation (c.1642G>T:p.Glu548*), which has been identified in 2 affected siblings, and variants in *POLE*, *POLD1* and the MMR genes. Although relatives were not always available for the patients found to carry variants in the latter group of genes, their established roles in inherited predisposition to colorectal tumours and availability of basic functional studies made it reasonable to investigate the pathogenicity of the variants.

The work in this part of the project required database interrogation to aid in appraising identified variants, and also communication with international experts. This highlights the importance of communication and collaboration in clinical research. It is apparent that genetic variants identified through clinical research projects are often not recorded in publicly accessible databases. This is a great shame. It would be excellent if *all* variants were recorded in a centralised repository, such as LOVD, as this would allow improved understanding of disease pathogenesis and therefore improved patient care. With the advent of the 100 000 Genomes Project it is hoped that there will be improved understanding of the role of the genome in health and disease. The project aims to sequence 100 000 genomes of individuals with cancer/ rare diseases/ infectious diseases. Genomes will be sequenced, annotated and analysed in the context of the individual's medical history. The project seeks to enhance the clinical interpretation of the data and derive new findings from the data (https://www.genomicsengland.co.uk/faqs-about-gecip/ accessed 10/5/2018). The Human Variome Project has a similar ethos. It is an international non-governmental organisation which functions to ensure that all information on genetic variation and its effects on human health can be collected, crated, interpreted and shared freely and openly (http://www.humanvariomeproject.org/about/about-the-human-variome-project.html accessed 10/05/18). The project itself doesn't physically store the data, but it develops standards and approaches so that data from different sources can be shared in an appropriate manner. LOVD is one example of a database which now functions under the auspices of the Human Variome Project.

4.4.2 Whole Exome Sequencing

24 patients underwent WES to try and identify further genes which may be implicated in colorectal neoplasia. Two approaches were taken: a candidate gene approach and a novel gene approach (Bradford 2017). Using the candidate gene approach, 106 variants were selected for validation: only 6 of these were not validated. In one of these cases, *ERCC6* in Halo05, there was a synonymous change at the base adjacent to that called, so it is likely that the bioinformatic analysis was miscalling the actual change. In 3 cases also occurring in Halo05, in the *ATM* gene, there was poor coverage at the loci, so the variants called are likely to be artefactual. In Halo15, a *ZKSCAN4* variant was not validated, but again there was low coverage at this locus, and the variant was called in only 2/3 reads. The final variant which was not validated was an *RBM5* frameshift in Halo27. The variant was only called in 5/22 reads, and it was at a region where there were numerous AC dinucleotide repeats, so the call was likely to be an artefact.

There are several important points to note about the WES results using the candidate gene approach:

- There were no obvious mutations which could explain the reduced *APC* expression for samples Halo52, Halo53 and Halo64.
- There was not a 'second hit' in a gene involved in BER for Halo40, who is a monoallelic carrier of a pathogenic *MUTYH* mutation.
- There were no mutations for Halo46 which were more likely to be pathogenic than her previously-identified *APC* mutation.
- In Halo47 and Halo68 there were no mutations in genes involved in the pathogenesis of oligodontia/ ectodermal dysplasia (*CXORF5, DLX1, DLX2, EDA, EDAR, EDARADD, FGFR1, GLI2, GLI3, LEF1, LTBP3, MSX1, NEMO*, *PAX9PITX2, P68, Wnt10a.* Chhabra *et al* 2014; Deshmukh *et al* 2012) (See Chapter 5).
- Of the truncating mutations, which are generally more likely to be clinically significant than missense changes:
    - The *BCLAF1* variant identified in Halo08 was not present in her affected son, whose DNA became available at a late stage in the project.
    - The *DSC2* variant in Halo17 may be implicated in a clinical phenotype other than colorectal neoplasia (cardiomyopathy) and no

relatives were available for screening, so it was not followed up in the short term as part of this project.

- A frameshift *POLQ* variant was identified in Halo18. I liaised with AmbryGen about this variant, and they commented that 'the subpopulation frequencies in 1K Genomes for POLQ c.4262_4268delTACTATT are high indicating that this would probably be likely benign based on our current classification scheme' (Felicia Hernandez personal communication 09/03/2017). Furthermore, relatives were not available for screening as the patient sadly passed away during the course of the project, so familial follow up was not possible.
- The *PTPRH* variant found in Halo27 may be involved in the pathogenesis of Parkinson's Disease (Appendix 4.9) and colorectal neoplasia is not typically part of the phenotypic spectrum, so it was not followed up in the short term as part of this project.
- The *RBMS2* variant in Halo28 had been mis-described in the bioinformatic analysis of the exome sequencing: it was actually a synonymous change rather than a stop-gain.
- The *BMP8B* variant in Halo55 was subsequently identified in his unaffected mother, as described above.
- The *RECQL* variant in Halo61 is interesting as *RECQL* is involved in DNA repair. Unfortunately, during the time course of this project it was not possible to recruit any relatives, but this variant will be pursued by the Inherited Tumour Syndromes Research Group, Cardiff University. Halo61 also carries a *POLE* mutation, which is being followed up in Chapter 5 of this thesis.

- The remaining variants, although all potentially interesting, were not followed up in the short term due to at least one of the following reasons:
  - No relatives available.
  - Variant more likely to be involved in a phenotype other than colorectal neoplasia.
  - Too common.
  - Insufficient evidence of a role in colorectal neoplasia to prioritise further study.

The second strategy in data analysis involved searching for novel polyposis genes (Bradford 2017). Bradford found one interesting variant. In Halo53, a novel *de*

novo stopgain in *MAP3K11* (p.Arg561*) was identified. *MAP3K11* is a mitogen activated protein kinase that is involved in the regulation of the c-Jun N-terminal kinase (JNK) pathway (Mishra *et al* 2007). There is some evidence that it may have a role in tumourigenesis: in HER+ breast cancer tissue, there is reduced function of MAP3K11 kinase activity (Das *et al* 2015). Furthermore, HER2-directed drugs such as trastuzumab and lapatinib, as well as depletion of HER2/ HER3, stimulate MAP3K11 kinase activity in HER2+ breast cancer cell lines (Das *et al* 2015). MAP3K11 has known pro-apoptotic effects and stable knockdown of MAP3K11 in the HER2+ cell line blunted the pro-apoptotic effects of trastuzumab and lapatinib. These findings suggest that HER2 activation inhibits the pro-apoptotic function of MAP3K11, which plays a mechanistic role in mediating anti-tumour activities of HER2-directed therapies (Das *et al* 2015). If the apoptotic effects of MAP3K11 are reduced, as may be the case with the p.Arg561* variant identified in Halo53, there may be enhanced tumour cell survival (Bradford 2017). Interestingly, there may be interactions between MAP3K11 and the Wnt-pathway. Although MAP3K11 can stabilise β-catenin, it has also been shown to inhibit conventional β-catenin/ TCF transcriptional activation (Thylur *et al* 2011). Therefore, MAP3K11 mutations might be expected to lead to enhanced expression of Wnt-target genes. The MAP3K11 variant is being further investigated by the Inherited Tumour Syndromes Research Group.

In this study, the most important factors limiting progress in the search for novel variants associated with polyposis were the lack of additional family members and the size of the study cohort. Other studies have been more successful because they were larger or prioritised the study of familial cases (Adam *et al* 2016; Weren *et al* 2015) with availability of samples from multiple affected relatives.

Although the mean depth of coverage across the 29 patients/ relatives who underwent WES was 61.99 reads, the percentage of the target region covered at 50x depth was only 43.9%. The coverage was also variable between patients, ranging from a mean of only 14.02 reads (Halo66) to a mean of 139.76 reads (Halo61). When apparent variants were not validated with Sanger sequencing, these tended to have occurred in regions of low coverage.

WES generates enormous amounts of data. Although a predetermined pipeline was used to analyse the data, the potential for human error remained. A literature search was performed for all the shortlisted variants to try and inform decisions

about which ones to pursue further, but the process remained subjective. This inevitably introduces bias into the process of data analysis. All the data is in the process of being re-analysed by a second, independent reviewer, Dr. Hannah West, but the scope for missing potentially significant variants will always remain. When analysing the results using the novel-gene approach, the volume of data potentially for literature review was significantly larger than when using the candidate-gene approach. In view of this, even more stringent criteria had to be applied before carrying out the literature search. Again, this introduces another layer of bias into data analysis. It is important to be aware that all pipelines used for WES data analysis included the criteria that the variant must have an allele frequency <0.5% in the general population. This would allow the identification of rare, dominant variants, but it means that recessive/ compound heterozygous diseases would only have been identified if they were very rare. This threshold was employed as it is comparable to the minor allele frequencies of pathogenic mutations identified in the recently identified autosomal recessive polyposis syndromes. In *NTHL1*-associated polyposis the mutant *NTHL1* variant has an allele frequency of 0.0036 in the control population (Weren *et al* 2015). In Adam's paper describing *MSH3*-associated polyposis the group had used a minor allele frequency of 1% as part of their filtering criteria for recessive variants, but the actual allele frequencies of the pathogenic mutations identified were 0.008%/ 0.0016%/ unreported (Adams *et al* 2016).

As anticipated, exome sequencing using both targeted and WES approaches identified many variants. These had to be prioritised in the short term, so variants were selected in genes which have an established role in inherited predisposition to colorectal tumours (*POLE*, *POLD1* and the MMR genes) or variants which were identified in families in which multiple family members were available for analysis (*AXIN2*). Assessment of variants using knowledge from databases is improving continuously. For example, in the 100 000 Genomes Project, different contributing centres may identify just 1 variant in a gene possibly associated with a particular phenotype. Once the data from all centres is collated, these variants can be linked so that an informative pattern emerges. Therefore review of the variants I have identified will be possible in the future, and may be helpful in highlighting further variants implicated in inherited polyposis.

4.5 Summary

This chapter utilised exome sequencing to try and identify novel genetic variants responsible for tumourigenesis in the cohort of NMI polyposis patients. Initially a targeted approach was used: 60 patients underwent deep sequencing of 15 genes known to be/ possibly involved in the pathogenesis of colorectal neoplasia. A subset of the cohort then had WES performed on germline DNA. A large number of variants were identified, from which variants in *AXIN2, POLE, POLD1* and the MMR genes were selected for further follow up in Chapter 5 of this thesis.

Chapter 5 Functional Characterisation of Variants Identified in the Mismatch Repair Genes, *POLE*, *POLD1* and *AXIN2*

5.1 Introduction

The final part of this thesis involved the genetic and functional characterisation of variants identified in Chapter 4. Mutations of interest were found in 3 genes/ groups of functionally similar genes: the MMR genes, two of the *pol* genes: *POLE* and *POLD1*, and *AXIN2*. This chapter is therefore divided into 3 sections, each addressing one of these genes/ groups.

5.2 The Mismatch Repair (MMR) Genes

5.2.1 Introduction

The human MMR system involves 7 key genes: *MSH2, MSH6, MSH3, MLH1, PMS1, PMS2* and *MLH3.* The role of five of these genes in CRC and Lynch Syndrome has been described in Chapter 1 (1.3 and 1.5.1.5).

As a result of targeted exome sequencing, three patients were found to carry potentially pathogenic variants in the MMR genes. They are summarised in Table 5.1:

Sample	Gene	Location	Variant	CADD Score	Amino Acid Change	Clinical Phenotype
Halo26	*MSH2*	2:47630458	A>G	26.2	p.Tyr43Cys	85-year-old female with 21 adenomas
Halo45	*MSH6*	2:48026228	C>T	29.5	p.Thr369Ile	77-year-old male with 27 adenomas and 7 HPPs
Halo70	*MSH6*	2:48030669	C>T	35	p.Arg1095Cys	71-year-old male with 15 adenomas

Table 5.1: Variants in the MMR genes identified through targeted exome sequencing

All three variants were investigated using a combination of database interrogation, MMR IHC and microsatellite instability testing.

5.2.2 Database Interrogation

The International Society for Gastrointestinal Hereditary Tumours (InSiGHT) is an international organisation which seeks to improve the quality of care for those individuals with any hereditary condition resulting in gastrointestinal tumours (https://www.InSiGHT-group.org/ accessed 31/08/2017). Part of their work includes the curation of a database which records variants in the MMR genes, along with a classification of the likelihood of pathogenicity. This database is recognised by clinicians as the 'gold standard' for the interpretation of such variants and is widely used in the clinical genetic diagnostic and research settings. There is also a French database, the Universal Mutation Database, UMD-MMR, which records variants in *MLH1*, *MSH2* and *MSH6* which have been reported in French laboratories in patients with CRC/ endometrial carcinoma (http://www.umd.be/ accessed 31/08/2017). It is also possible to calculate the prior probability of pathogenicity of

MMR gene variants, based on sequence conservation and position (Thompson *et al* 2013).
All these tools were employed in the investigation of the MMR variants which had been identified in NMI patients.

5.2.3 MMR Immunohistochemistry (IHC)

It is well established that mutations in the MMR genes, *MLH1, MSH2, MSH6* and *PMS2* lead to Lynch Syndrome (LS), which accounts for around 5% of cases of CRC (reviewed in Mishra and Hall 2012). LS-related cancers are characterised by the absence of MMR protein expression (reviewed in South *et al* 2009), and MMR IHC is a fundamental component of the molecular diagnosis of LS. Previous authors have reported that colorectal polyps which develop in carriers of known LS mutations show loss of MMR staining: this was observed in 79% of adenomas and 27% of serrated lesions (Walsh *et al* 2012).

I hypothesised that if the MMR mutations identified in NMI patients were pathogenic, the majority of adenomas in the patients would show loss of MMR protein staining.

5.2.4 Microsatellite Instability (MSI) Testing

Microsatellite instability (MSI) refers to a hypermutable phenotype caused by defective MMR activity (Boland and Goel 2010). It is the characteristic genetic signature identified in cancers associated with LS. It has previously been shown that 80% of adenomas from patients with LS show evidence of MSI (Iino *et al* 2000).

I hypothesised that if the MMR variants identified in the NMI patients were pathogenic, their adenomas would show evidence of MSI.

5.2.5 Methods

5.2.5.1 Database Interrogation

All variants were investigated to determine what their InSiGHT and UMD classifications were, and what their prior probability of pathogenicity was (Thompson *et al* 2013). The databases accessed are available at:

1. http://www.InSiGHT-database.org/genes
2. http://www.umd.be/MSH2/
3. http://www.umd.be/MSH6/
4. http://hci-lovd.hci.utah.edu/variants.php?select_db=MSH2_priors&action=search_all&search_Variant%2FDNA=c.128A%3EG, http://hci-lovd.hci.utah.edu/variants.php?select_db=MSH6_priors&action=search_unique&order=Variant%2FDNA%2CASC&hide_col=&show_col=&limit=100&search_Variant%2FExon=&search_Variant%2FDNA=&search_Variant%2FRNA=&search_Variant%2FProtein=p.T369I&search_Variant%2FCustom_PP2_score=&search_Variant%2FMAPP_score=&search_Variant%2FMAPP%2FPP2_Prior=&search_Variant%2FReference=&search_Variant%2FDetection%2FTemplate=&search_Variant%2FDetection%2FTechnique=&search_Variant%2FDBID=
5. http://hci-lovd.hci.utah.edu/variants.php?select_db=MSH6_priors&action=search_unique&order=Variant%2FDNA%2CASC&hide_col=&show_col=&limit=100&search_Variant%2FExon=&search_Variant%2FDNA=&search_Variant%2FRNA=&search_Variant%2FProtein=p.R1095C&search_Variant%2FCustom_PP2_score=&search_Variant%2FMAPP_score=&search_Variant%2FMAPP%2FPP2_Prior=&search_Variant%2FReference=&search_Variant%2FDetection%2FTemplate=&search_Variant%2FDetection%2FTechnique=&search_Variant%2FDBID=

(accessed 26/06/2017)

5.2.5.2 MMR IHC

FFPE tissue (multiple adenomas and HPPs) was available for all 3 patients in whom we had identified MMR gene variants: Halo26 (6 x TVA LGD, 7 x TA LGD, 1 x VA LGD), Halo45 (13 x TVA LGD, 2 x HPP) and Halo70 (2 x TVA LGD, 13 x TA LGD, 1

x HPP). All FFPE tissue blocks underwent IHC analysis of MLH1, MSH2, MSH6 and PMS2 at the Cellular Pathology Department, University Hospital of Wales. The EnVision FLEX System (Agilent) was used, following the manufacturer's protocol, and the primary antibodies MLH1 Flex RTU, MSH2 FE11, MSH6 EP49 and DAKO PMS2 1/80. I reviewed the slides using an Olympus BX43 light microscope.

5.2.5.3 MSI Testing

DNA was extracted from FFPE polyps by the AWMGS using the Maxwell Promega LEV FFPE kit. MSI testing was carried out by the AWMGS following the protocol given in Appendix 5.1.

5.2.6 Results

5.2.6.1 Database Interrogation

The results from the database interrogation are summarised in Table 5.2 (InSiGHT class 3: uncertain; UV: uncertain variant):

Gene	Chr	Position	Variant	Amino Acid Change	InSiGHT/ UMD	Prior Probability of Pathogenicity (by method of Thompson *et al* 2013)
MSH2	chr2	47630458	A>G	p.Tyr43Cys	InSiGHT class 3; 4 reports on UMD - UV	0.8980
MSH6	chr2	48026228	C>T	p.Thr369Ile	Not reported	0.0462
MSH6	chr2	48030669	C>T	p.Arg1095Cys	Not reported on InSiGHT; 5 reports on UMD – UV	0.9342

Table 5.2: Results of MMR gene database interrogation

The *MSH2* variant, c.128A>G:p.Tyr43Cys, has an InSiGHT classification of 3, meaning that its pathogenicity is uncertain. The prior probability of pathogenicity is 0.8980, in concordance with its InSiGHT classification. The *MSH6* variant, c.1106C>T:p.Thr369Ile, has not previously been recorded in the InSiGHT/ UMD databases. However, it has a low prior probability of pathogenicity score: 0.0462. This would put it in the InSiGHT class 2 category (https://www.InSiGHT-group.org/criteria/) i.e. likely not pathogenic. The variant in Halo70, *MSH6* c.3283C>T:p.Arg1095Cys, had not previously been reported in the InSiGHT database. There were 5 reports on UMD, and it was recorded as a UV. It had a relatively high prior probability of pathogenicity score: 0.9342. This would place it in the InSiGHT class 3 category (https://www.InSiGHT-group.org/criteria/).

5.2.6.2 MMR IHC

Attempted IHC staining of the MMR proteins was performed on all tumours from the 3 patients. One TA LGD from Halo45 had cut out, so there was no tissue available

for MMR IHC, and there was insufficient tissue available for complete MMR IHC for one HPP from Halo45, so results for MLH1 were not available.

For the tumours with successful MMR IHC, there was no evidence of loss of MMR staining in any of the lesions and the staining pattern was normal.
A representative image of the MMR IHC is shown in Figure 5.1:

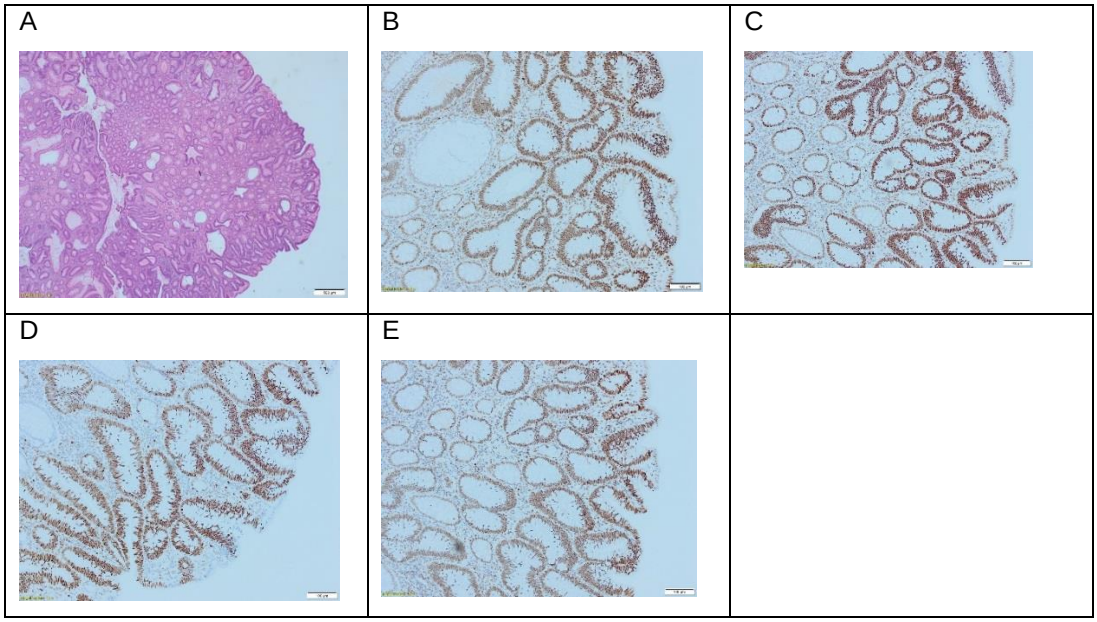


Figure 5.1: Halo70 TA LGD. A: H&E, B: MLH1 IHC, C: MSH2 IHC, D: MSH6, E: PMS2

5.2.6.3 MSI Testing

A subset of 20 tumours underwent MSI testing (Appendix 5.2). The remaining tumour samples did not undergo screening due to significant diathermy artefact/ the lesion cutting out/ a small number of dysplastic glands being present amongst normal glands/ the specimen being too small.

There was no evidence of MSI in any of the tumours tested.

5.2.7 Conclusions of MMR Gene Investigations

Database interrogation, MMR IHC and MSI testing was performed for the patients in whom variants in the MMR genes had been identified through targeted exome sequencing. Database interrogation showed that all three variants were likely to be

of uncertain pathogenicity (*MSH2* c.128A>G:p.Tyr43Cys and *MSH6* c.3283C>T:p.Arg1095Cys) or were likely not pathogenic (c.1106C>T:p.Thr369Ile). Tumour samples from the carriers of these variants showed normal MMR IHC and no evidence of microsatellite instability. Although relatively small numbers of tumours underwent testing, it is unlikely that any of the identified mutations are clinically significant.

There were no relatives available for any of the patients to be screened.

In summary, 3 patients were identified as carrying variants in the MMR genes. Although polyposis is not the typical phenotype seen in heterozygous carriers of MMR gene mutations, these variants were investigated to gain a greater understanding of whether they might have significant functional effects. It is unlikely that any of the variants are clinically significant, in view of the normal MMR IHC and microsatellite stability of the tumours.

5.3 *POLE* and *POLD1*

5.3.1 Introduction

Mutations in the proofreading domains of the human polymerase enzymes, *POLE* and *POLD1*, have been shown to have a role in the pathogenesis of colorectal neoplasia, as described in Chapter 1 (1.5.1.3). As a result of the targeted exome sequencing, four patients were found to carry missense variants in *POLE/ POLD1.* They are summarised in Table 5.3, along with the patients' phenotypes and the location of the resultant amino acid change in the protein:

Patient ID	Gene	Posn.	Variant	CADD Score	Amino Acid Change	Location of Amino Acid Change	AmbryGen Class	Patient Phenotype
Halo15	*POLE*	12:133249812	T>C	24.4	p.Met471Val	In ribonuclease H-like domain, DNA directed DNA polymerase Family B exonuclease domain	Not reported	84-year-old male with 13 TA LGD
Halo18	*POLE*	12:133202816	C>T	23.8	p.Glu2140Lys	No domain	VUS	72-year-old male with 27 polyps, majority TAs
Halo27	*POLD1*	19:50919693	C>T	34	p.Thr954Met	In DNA directed DNA polymerase family B multifunctional domain	Not reported	67-year-old female with 17 adenomas
Halo61	*POLE*	12:133245452	T>C	26	p.Tyr623Cys	In DNA directed DNA polymerase family B multifunctional domain	VUS	60-year-old female with 11 polyps: 2 SAs, 7 TA, 1 TVA, 1 HPP

Table 5.3: Variants in the pol genes identified through targeted exome sequencing. The amino acid location was determined using http://www.ncbi.nlm.nih.gov/Structure/cdd/wrpsb.cgi?INPUT_TYPE=live&SEQUENCE=NP_006222.2 (accessed 09/03/16)

Only one of the variants was located in the exonuclease domain of the protein, POLE p.Met471Val in Halo15. All of the pathogenic variants that have been reported in the literature thus far are also located within the exonuclease domain of POLE/ POLD1 (Palles *et al* 2013). However, I hypothesised that a variant may be pathogenic, even if it lies outside of the exonuclease domain, if it can alter the ability of the exonuclease domain to elicit its functional effects, for example through altering the movement of DNA through the active site of the polymerase.

In Palles' paper (Palles *et al* 2013), the group examined 39 tumours from 11 *POLE* p.Leu424Val carriers for somatic mutations. Second hits by LOH involving the germline wildtype allele were found in some tumours and all tumours were microsatellite stable. Most tumours were screened for *KRAS* and *BRAF* driver mutations, and a sub-set of tumours were screened for known pathogenic mutations in *APC, CTNNB1, PIK3CA* and *FBXW7*. Mutations were all base substitutions. This was especially interesting for *APC,* as ~60% of mutations seen in sporadic tumours are frameshifts (reviewed in Palles *et al* 2013). In addition to this, certain sites not commonly mutated in sporadic colorectal tumours seemed to be mutation hotspots, for example codons 1114 and 1338 of *APC* and codon 146 of *KRAS*. In view of this, I hypothesised that the same 'hot spot' locations may be mutated in tumours obtained from the patients we identified as harbouring germline *POLE* variants, if the variants were pathogenic. I also hypothesised that the tumours should be microsatellite stable, in keeping with Palle's results.

Palle's group is currently carrying out work to determine whether tumours developing in patients carrying pathogenic mutations in the *pol* genes have a characteristic mutational signature. We collaborated with their group to examine the somatic mutations in tumours occurring in the patients we identified as carrying *pol* gene variants.

The genetic variants identified were therefore investigated through a combination of:

- Modelling the mutant proteins to investigate whether the amino acid changes could theoretically affect movement of DNA through the exonuclease domains of the polymerases. This work was carried out in collaboration with Dr. Pierre Rizkallah, Cardiff University
- Somatic mutation screening, to identify 'hotspot' mutations as described by Palles *et al* (2013)

- MMR IHC and MSI testing to determine whether tumours were microsatellite stable
- Segregation/ *de novo* analysis in families, where family members were available for testing
- Investigation of the mutation signature of the tumours. This work was carried in collaboration with Professor Ian Tomlinson and Dr. Claire Palles (Oxford University/ Birmingham University)

5.3.2 Methods

5.3.2.1 *In Silico* Modelling of *Pol* Variants

Thank you to Dr. Pierre Rizkallah for his expert help with this part of the project.

The Protein Databank (PDB) (http://www.rcsb.org) server's BLAST based sequence alignment algorithm was used to match the human POLD1 sequence with the sequence of many POLD1 entries represented in the structural database. The top hit was **3IAY** (Swan *et al* 2009), with the alignment shown in Appendix 5.3a. The sequence of the human POLE had the top hit **4M8O** (http://www.rcsb.org, Hogg *et al* 2014). The alignment is shown in Appendix 5.3b. These models were then used to analyse the effect of variants on a per residue basis. The program Crystallographic Object-Oriented Toolkit (COOT) (Emsley and Cowtan 2004) introduced the identified mutations and regularised the geometry in the immediate neighbourhood. PyMOL (Delano 2002) was used to produce graphical images to visualise the results of the predictions.

5.3.2.2 Somatic 'Hot Spot' Mutation Screening in *APC* and *KRAS*

At the time of the experiments, a total of 15 FFPE tumours were available for Halo15 and Halo61 (Halo15: 11 x TA LGD; Halo61: 4 x TA LGD) (Appendix 5.4). A further 45 tumours were obtained from patients with variants in *POLE* that were assumed to be non-pathogenic, due to their AmbryGen classification (4.1.4.2.4)

(Halo31: 10 x TA LGD, 2 x HPP; Halo69: 6 x TA LGD, 10 x HPP, 1 x CRC; Halo81: 3 x TA LGD, 3 x TVA LGD, 9 x HPP, 1 x TVA arising in a SSL) (Appendix 5.4). These were included in the study protocol to ensure that somatic 'hotspot' mutations were not present: if such mutations had been identified it would suggest that the designation of the variants as benign may not be correct. DNA was extracted from the tumours using the GeneRead DNA FFPE Kit (Qiagen) following the standard protocol (2.3.4). DNA underwent PCR and Sanger sequencing using standard reagents and reaction conditions for DNA extracted from FFPE tissue (2.3.8.2, 2.3.8.5, 2.3.9-2.3.13). Tumours were screened at the 'hotspot' mutation locations in *APC* and *KRAS* described by Palles *et al* (2013): *APC* codons 1114 and 1338 and *KRAS* codon 146. Primers were designed using primer3 software and were supplied by Eurofins (Appendix 5.5).

5.3.2.3 MMR IHC and Microsatellite Stability of Tumours in Carriers of *Pol* Gene Variants

34 FFPE tumours from Halo15, Halo27 and Halo61 were available (Halo15: 11 x TA LGD; Halo27: 17 x TA LGD, 1 x SSL; Halo61: 5 x TA LGD (Appendix 5.6).
All FFPE tumours underwent IHC analysis of MLH1, MSH2, MSH6 and PMS2 at the Cellular Pathology Department, University Hospital of Wales. The EnVision FLEX System (Agilent) was used, following the manufacturer's protocol, and the primary antibodies MLH1 Flex RTU, MSH2 FE11, MSH6 EP49 and DAKO PMS2 1/80.
A subset of 28 tumours underwent DNA extraction. This was done by the AWMGS service using the Maxwell Promega LEV FFPE kit. MSI testing was carried out following the protocol given in Appendix 5.1. The remaining tumours did not undergo MSI testing due to the dysplastic glands being too scanty to warrant DNA extraction or there being significant diathermy artefact.

5.3.2.4 Segregation Analysis

I sought to recruit relatives of the patients in whom we had identified germline variants in *POLE* and *POLD1* to try and determine whether the variants were occurring *de novo* in the index cases or whether they were segregating with disease if multiple family members were affected.

Halo15 was an only child and there was no family history of colorectal neoplasia. Halo18's mother and maternal aunt had a possible history of CRC, reported by the family. Halo18 has two siblings who are thought to be unaffected. Halo27 has 3 siblings with colorectal polyposis and her father had CRC. Halo61 has one sibling with colorectal polyposis and 2 unaffected siblings. She had a son who died of CRC at the age of 23. The son's tumour IHC was suggestive of a MMR defect, but there was insufficient DNA for MSI testing. Halo61 also has a daughter who had developed a single colorectal polyp at the age of 28. There was, in addition, a history of CRC on her husband's side of the family.

The only relatives that were successfully recruited to the study were both children of Halo15. The daughter of Halo15 was a 43-year old female. She had a colonoscopy at age 30 for rectal bleeding: it showed colonic endometriosis but no polyps. The son was a 42-year old male who has never had a colonoscopy. Germline DNA was extracted from whole blood by the AWMGS (2.3.2). RNA was extracted from whole blood following the standard protocol (2.3.3). RNA was reverse transcribed to cDNA (2.3.3.2). DNA and RNA underwent PCR and Sanger sequencing to screen for the *POLE* mutation, c.1411A>G:p.Met471Val. Standard reagents and reaction conditions were used (2.3.8.1, 2.3.8.5, 2.3.9-2.3.13). DNA primers were those used for Haloplex validation (Appendix 4.2). cDNA primers are listed in Appendix 5.7.

5.3.2.5 Investigation into the Mutation Signature of Tumours

This work was carried out with Professor Ian Tomlinson's group at the Wellcome Trust Centre for Human Genetics, Oxford University. I am extremely grateful for their help and involvement.

The mutation signature of tumours occurring in patients with germline *POLE/ POLD1* variants was determined by examining 30 genes. For some genes, the full coding regions, with 10 bp flanking region were interrogated. For others, hotspots or domains were selected. The genes interrogated were: *ACVR2A, APC, ARID1A, ATM, B2M, BCL9L, BMPR2, BRAF, CTNNB1, ELF3, FBXW7, GNAS, KRAS, MLH1, MSH2, MSH6, NRAS, PIK3CA, POLD1, POLE, PTEN, RNF43, RPL22, SMAD2, SMAD4, SOX9, TCF7L2, TGIF1, TP53, ZFP36L2.*

The genes which didn't have the full coding region examined were: *APC* (codons 1-1600), *POLD1* (chr19:50905938-50910304), *POLE* (chr12:133249809-133253238), *GNAS* (codon 200), *BRAF* (exons 11 and 15: chr7: 140453074-140453193, chr7:140481375-140481493), *CTNNB1* (hotspots chr3:41266444-41266698, chr3:41267150-41267352, chr3:41274831-41274935, chr3:41275019-41275358, chr3:41277214-41277334, chr3:41266016-41266244), *PIK3CA* (hotspots chr3:178916613-178916965, chr3:178935997-178936122, chr3:178951881-178952152).

Gene capture was achieved using single molecule molecular inversion probes (smMIPs), which were designed by Dr. Palles using MIPgen software and were supplied by Integrated DNA Technologies. A total of 300 smMIPs were used to target 63.1kb. The smMIPs capture ~1800 sites which are commonly mutated when there is impaired POLD1/ POLE proofreading due to a functional mutation in the exonuclease domain (Palles personal communication).

I travelled to Oxford to assist with the smMIP gene capture. The protocol for the capture and subsequent sequencing is described in Chapter 2 (2.4.6).

5.3.3 Results

5.3.3.1 Modelling of *Pol* Variants

The 3-dimensional location of the variants is described and shown, along with their potential functional effects, in Table 5.4:

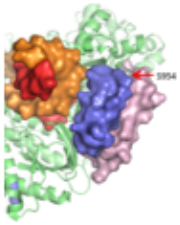
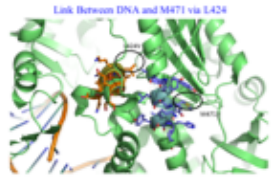
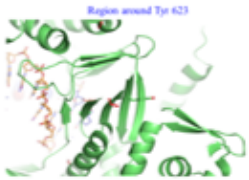
Patient	Gene	Amino Acid Change	Information from Protein Structure	Image
Halo27	*POLD1*	p.Thr954Met	The amino acid change is likely to distort the structure of the active site of the exonuclease domain, so may have pathogenic effects.	
Halo15	*POLE*	p.Met471Val	The mutation is in the exonuclease domain in the vicinity of the recurrent pathogenic L424V variant.	
Halo18	*POLE*	p.Glu2140Lys	Not in model	N/A
Halo61	*POLE*	p.Tyr623Cys	The amino acid change is predicted to result in an unpaired cysteine, which could bind to nearby cysteines and potentially distort the protein structure in the vicinity of DNA as it moves through the active site of the protein.	

Table 5.4: 3-dimensional location of variants in POLE/ POLD1 and predicted functional consequences

5.3.3.2 Somatic 'Hot Spot' Mutation Screening in *APC* and *KRAS*

60 tumours from patients carrying presumed benign or putatively pathogenic *Pol* gene variants were screened for mutations at *APC* codons 1114 and 1338, and *KRAS* codon 146. No mutations were identified.

5.3.3.3 MMR IHC and Microsatellite Stability of Tumours in Carriers of *Pol* Gene Variants

34 tumours from patients carrying putatively pathogenic variants in *POLE/ POLD1* were screened using MMR IHC +/- MSI testing. There was no evidence of loss of MMR protein staining or MSI in any of the samples (Appendix 5.8).

5.3.3.4 Segregation Analysis

The son and daughter of Halo15 both underwent screening for the *POLE* variant, c.1411A>G:p.Met471Val. The variant was present in the daughter only, in both DNA and cDNA (sequencing traces in Appendix 5.9).

5.3.3.5 Investigation into the Mutation Signature of Tumours

A total of 2 tumour samples from Halo15 have thus far undergone successful smMIP gene capture and sequencing. Both lesions were TA LGD. Tumour samples from Halo27 are currently being analysed: DNA from 7 x TA LGD has been submitted. The tumour DNA (one TA LGD) from Halo61 failed quality control and no results are available.

At the time of writing, the high quality raw results for somatic mutation sequencing for Halo15 were available (Appendix 5.10). However, these were being further analysed by the bioinformatics team at Birmingham University, to determine whether they are consistent with a hypermutated signature observed in tumours occurring in carriers of known pathogenic mutations in the *pol* genes.

5.3.4 Conclusions of *POLE/ POLD1* Gene Interrogation

Four patients were found to carry potentially pathogenic germline mutations in *POLD1* or *POLE.* All variants were missense changes. The variants were investigated using a combination of database interrogation, modelling the variants using *in silico* structural modelling, MMR IHC and MSI testing, somatic mutation analysis and family segregation studies.

Based on *in silico* modelling, two of the variants were thought most likely to be disease-causing. One was a POLD1 variant in Halo27, p.Thr954Met, which was predicted to distort the structure of the active site of the enzyme. It had a very high CADD score of 34. The second variant was identified in Halo15, p.Met471Val. This is a novel variant which is located in the exonuclease domain of the protein.

The remaining two variants included a POLE variant in Halo18, p.Glu2140Lys, and a POLE variant in Halo61, p.Thr623Cys. POLE p.Glu2140Lys is reported as benign on dbSNP. It does not lie within any of the POLE functional domains and was not in the protein model used as part of this project. The variant is considered very unlikely to be clinically significant.

POLE p.Thr623Cys was suggested to possibly have pathogenic effects that could be mediated through an alteration in the protein structure in the proximity of DNA as it moves through the active site of the enzyme (Dr. Pierre Rizkallah, personal communication). It has a CADD score of 26 and is classified as a VUS by AmbryGen. Tumour DNA from Halo61 did not have any evidence of 'hot spot' mutations in *APC* or *KRAS*, although only 4 lesions were available for analysis. Unfortunately, tumour DNA was not successfully sequenced following smMIP gene capture, so its mutation signature remains uncertain. Therefore, at the present time, there is insufficient evidence to support or refute a pathogenic effect of this mutation and it remains a VUS.

Tumours from Halo15 and Halo61 underwent testing for somatic 'hotspot mutations' described by Palles *et al* (2013). No such mutations were identified, however this is not surprising: in Palles' paper, the mutations were present at the following frequencies:

Variant	Prevalence in Adenomas	Prevalence in Carcinomas
APC p.Arg1114* (C>T)	1/36 (3%)	1/3 (33%)
APC p.Gln1338* (C>T)	3/36 (8%)	1/3 (33%)
KRAS p.Ala146Thr (A>C)	1/36 (3%)	1/3 (33%)

There were 11 tumours available for Halo15/ p.Met471Val and 4 for Halo61/ p.Tyr623Cys. This study was therefore underpowered to identify mutations at the hotspot locations, as we had insufficient numbers of tumours. Additionally the dysplastic epithelial component of the lesions was highly variable – ranging from 10%-100%. For the lesions with a small dysplastic component, even if a hotspot mutation was present it may not have been observed with Sanger sequencing.

Tumour material for Halo27 only became available at a late stage of the project. Tumour DNA did not undergo screening for the 'hot spot' variants but is undergoing mutation signature analysis, which potentially has a greater power to support or refute the pathogenicity of the *POLD1* variant.

All of the tumours from Halo15, Halo27 and Halo61 which underwent MSI testing were found to be microsatellite stable.

The only relatives available for family studies were the son and daughter of Halo15. The daughter was found to carry the same variant as her affected father, and this was confirmed to be present in both DNA and cDNA. However, her phenotype is not accurately known, as she has not had a recent colonoscopy. The genetic finding has been fed back to her clinician, who has arranged surveillance colonoscopy.

SmMIP gene capture is being performed on tumour DNA from Halo15 and Halo27, in collaboration with Professor Ian Tomlinson's group at Oxford University/ Birmingham University. The aim of this work is to determine whether the tumours from these patients have a hypermutated genetic signature, in keeping with that observed in the tumours which occur in carriers of known pathogenic mutations in

the *pol* genes. Although the raw results are available for 2 TA LGD from Halo15, bioinformatic analysis is still pending. There was a seven-month delay between carrying out the sample preparation and obtaining the raw data, due to factors outside of our control. It is anticipated that the complete analysis will be finalised within the next several months. Therefore, at present, the variants identified in Halo15 and Halo27, c.1411A>G:p.Met471Val and c.2861C>T:p.Thr954Met respectively, remain as VUS.

5.4 *AXIN2* and *AXIN2*-Associated Polyposis (AxAP)

5.4.1 Introduction

The role of the Wnt-pathway in colorectal neoplasia has been described in the previous chapters of this thesis. AXIN2 is a scaffold protein which is involved in regulating Wnt signalling. It forms part of the 'destruction complex' which targets β-catenin for degradation: AXIN2 supports the GSK3β-dependent phosphorylation of β-catenin, which marks the protein for ubiquitination and subsequent proteosomal degradation (reviewed in Wu *et al* 2012). The AXIN2 protein has several domains, including a Tankyrase binding domain which regulates protein stability, an RGS domain which mediates binding to APC, a β-catenin binding domain and a DIX dimerization domain (reviewed in Mazzoni and Fearon 2014). AXIN2 is a transcriptional target of β-catenin dependent Wnt signalling, and its levels are elevated in cancers with Wnt activating mutations, therefore potentially negatively regulating Wnt signalling (reviewed in Mazzoni and Fearon 2014).

In view of its effects on Wnt signalling, it is possible that germline *AXIN2* mutations might be associated with an increased risk of colorectal neoplasia. In 2004, Lammi *et al* described a Finnish family with oligodontia (Lammi *et al* 2004). The family exhibited oligodontia segregating as an AD trait across 4-generations. Two patients from the oldest generation were known to have a history of colorectal neoplasia (one had metastatic CRC and the other had 68 adenomas) and 10 further family members underwent a colonoscopy/ sigmoidoscopy: 7 with oligodontia and 3 healthy individuals. Colorectal neoplasia was identified in 6 individuals with oligodontia, but not in the unaffected individuals. The colorectal phenotype was highly variable, ranging from a single HPP to 10-20 HPP plus 2 adenomas. Family members were screened for *AXIN2* mutations in coding regions/ flanking intronic sequences, and a c.1966C>T transition in exon 7 of was identified. This results in a p.Arg656* codon and premature termination of translation. The mutation was present in 11 individuals with oligodontia, but in none of 6 healthy family members. The group also identified a further *AXIN2* truncating mutation in a 13-year-old boy with oligodontia. They describe a 1bp G insertion after nucleotide 1994 in exon 7, which results in re-coding of the amino acids starting at p.Asn666 and a stop codon, 40 codons later. The same mutation had previously been identified as a somatic mutation in CRC (reviewed in Lammi *et al* 2004).

Lammi *et al* concluded that their results provide strong evidence that familial colorectal cancer can be caused by mutations in *AXIN2*.

In 2011, Marvin *et al* reported the identification of a novel *AXIN2* mutation, c.1989G>A, in a family with a history of oligodontia, absent eyebrows, sparse hair, colonic polyps, early onset CRC and early onset breast cancer (Marvin *et al* 2011). The colorectal phenotype included an individual with >100 adenomas and an individual with 2 metanchronous CRC plus 5 adenomas. The mutation introduces a stop codon at amino acid 663, p.Tyr663*, in exon 7 (Marvin *et al* 2011). It was identified in 3 individuals with a phenotype of oligodontia and colorectal neoplasia, and was absent in 2 unaffected family members. Subsequent *in vitro* transcription and translation of a c.1989G>A construct and expression of the construct in HEK293T cells produced a truncated AXIN2 product (Marvin *et al* 2011). The group commented that their findings provided further evidence of an AD multisystem ectodermal and neoplastic phenotype associated with a germline *AXIN2* mutation. They noted that the truncated protein would lack its DIX dimerization domain, which they predicted would impair the inhibitory action of AXIN2 on WNT signalling.

Functional studies subsequently carried out on the mutation identified by Marvin *et al* (2011), *AXIN2* c.1989G>A, revealed that the mutation did not result in nonsense mediated decay (NMD), and that it could cause activation of the Wnt pathway, although this was context dependent (Mazzoni *et al* 2015). The group transiently transfected HEK293T cells with wild-type and mutant-containing plasmids. They showed that truncated AXIN2 protein was more abundant than wild-type (WT) protein, despite equivalent transcript levels. Interestingly, despite lacking the dimerization domain, truncated AXIN2 was still found to interact with both WT-AXIN2 and truncated AXIN2. It was suggested that such interactions may not be direct but might be mediated by another protein in the destruction complex, such as APC, which contains AXIN binding sites. Although truncated AXIN2 was able to interact with AXIN2, its ability to bind to AXIN1 was impaired. An important point to note is that when overexpressed, truncated AXIN2 was able to inhibit SW480 CRC cell colony formation and was able to inhibit β-catenin/ T cell factor- dependent reporter gene activity, as would be expected with WT protein. However, when stably expressed in rat intestinal IEC-6 cells, truncated AXIN2 did not match WT AXIN2's ability to inhibit Wnt-mediated expression of Wnt-regulated target genes (Mazzoni *et al* 2015). It was suggested that when AXIN2 levels are low, such as in the absence of Wnt-stimulation, the mutation may have a loss-of-function effect.

When there is elevated *AXIN2* expression, the truncated protein may have gain-of-function effects.

In 2014, Rivera *et al* identified a further novel variant in *AXIN2* in a 3-generation family with a phenotype of multiple colorectal adenomas and/ or CRC, but without any signs of ectodermal dysplasia/ oligodontia. The missense variant, c.1387C>T, p.Arg463Cys is located in exon 5 within the β-catenin binding domain. However, its pathogenicity is questionable, as the mutant allele was apparently lost in polyps from the proband and his sister (Mazzoni and Fearon 2014), and the mutation was found to be present in an unaffected relative (Rivera *et al* 2014).

Although these three papers have identified potentially pathogenic germline *AXIN2* mutations in patients with colorectal polyposis/ CRC, there are also several studies which have failed to identify pathogenic variants, although the study cohorts were small (Mongin *et al* 2012; Lejeune *et al* 2006). Mongin *et al* (2012) did not identify any mutations in the coding sequence of *AXIN2* in a cohort of 38 NMI patients with >40 adenomas or >20 polyps, and Lejeune *et al* (2006) failed to identify pathogenic variants in *AXIN2* in a cohort of 31 patients with multiple colorectal adenomas (18 individuals), one of whom also had tooth agenesis, or microsatellite stable CRC (13 individuals). However, in the latter paper, the patients had only undergone *APC* screening to define their NMI status, and when *MUTYH* was analysed as part of the study protocol, 32% of the patients were found to have MAP.

Chapter 4 described the targeted exome capture and UDS of genes which might be implicated in polyposis/ CRC. A sibling pair was found to carry a truncating mutation in *AXIN2:* c.1642G>T:p.Glu548* (Halo47 and Halo68, Family A). This section of Chapter 5 will focus on this mutation and approaches to investigate its possible pathogenicity.

5.4.2 Halo47 and Halo68: Family History and Clinical information

Halo47 was a 51-year-old female when she was recruited to the study. In 2013, at the age of 50, she had removal of 2 TVAs with LGD and 3 TAs with LGD. A subsequent colonoscopy found >50 sessile polyps, up to 10mm, throughout the colon. She underwent a subtotal colectomy, and approximately 40 sessile polyps

were identified macroscopically. Four blocks were taken of polyps: the majority were TA LGD, but 2 were serrated adenomas (it is not clear exactly how many polyps were sampled). The patient was due to have a completion colectomy, but at the time of surgery she was found to have a large rectal cancer. She underwent sigmoidectomy/ proctectomy/ abdominoperineal resection and the cancer was reported as being pT4bN2MX, Duke's C1. Unfortunately, she developed liver and lung metastases and she passed away during the course of this project.

Her sister, Halo68, had a colectomy for colorectal polyposis at the age of 50. She had 14 TA with LGD and 6 HPPs.

Both Halo47 and Halo68 had been diagnosed with colorectal neoplasia following symptomatic presentation.

There is a family history of intestinal neoplasia. Their mother (1.2) had CRC (moderately differentiated Duke's A adenocarcinoma), and her medical records described a VA with LGD and 'three small metaplastic polyps' identified macroscopically, one of which was confirmed microscopically to be a HPP. When all available specimens were reviewed for this research, she was found to have had at least one VA LGD, CRC arising in a VA, one HPP and one TA LGD.

The sisters Halo47 and Halo68 had a maternal aunt (1.1) with pT3N2MX CRC and 8 polyps (5 were sampled and were found to be TA with LGD) and a maternal aunt with 'stomach cancer' reported by the family (it was not possible to access her medical records).

The family tree is illustrated in Figure 5.2 and the family is hereafter referred to as Family A:

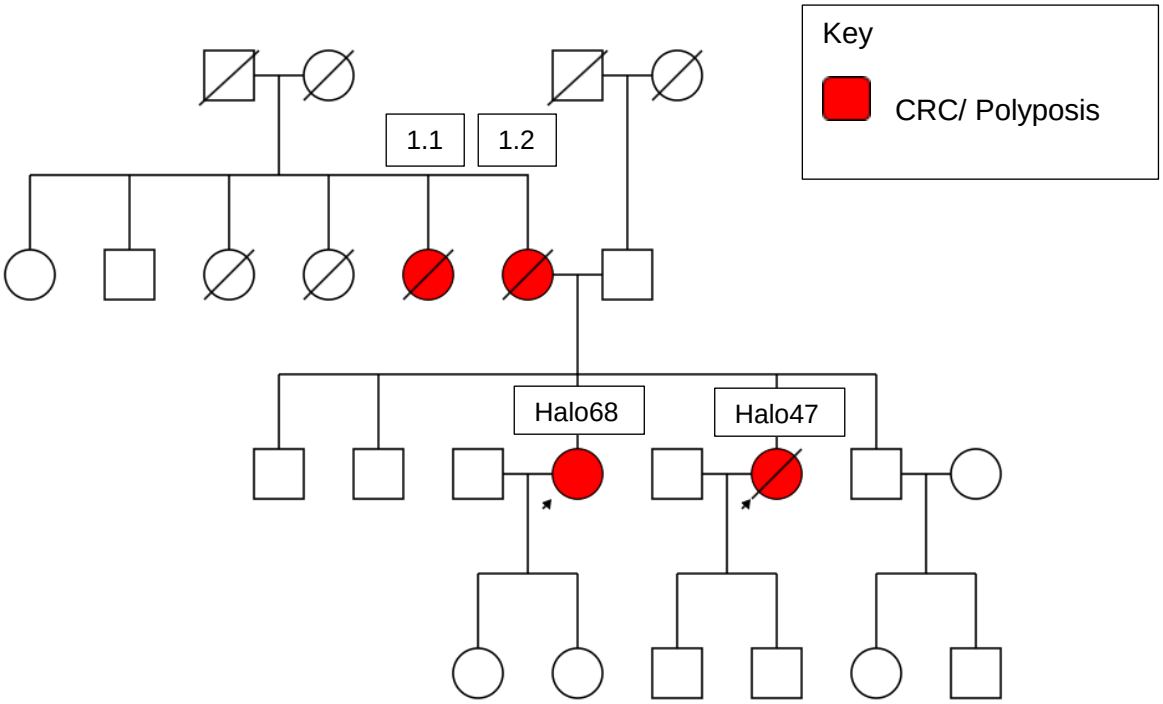


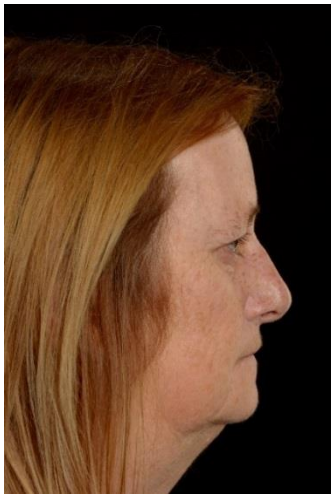
Figure 5.2: Family tree of Family A

Both sisters and their mother allegedly lacked all adult teeth. This was confirmed in Halo68 by a formal dental assessment at the University Hospital of Wales dental department. Halo68 was reviewed in clinic and was found to have sparse hair on the outer third of her eyebrows, thin hair on her head, and she was born with a malformation of the nail on her left index finger.

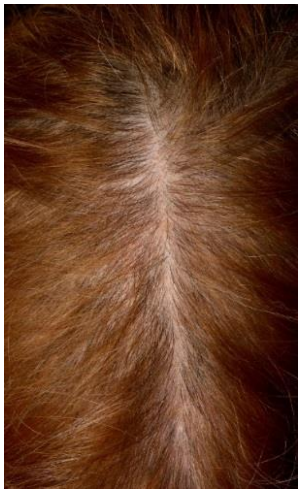
One of the daughters of Halo68 also has sparse hair on the outer third of her eyebrows, but her head hair and teeth are normal. Family photos of Halo47 and Individual 1.2 are suggestive of scanty eyebrow hair. See Figure 5.3.

A

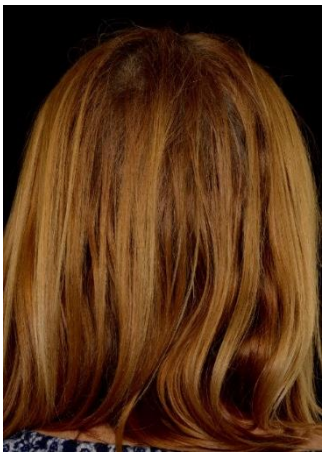
B



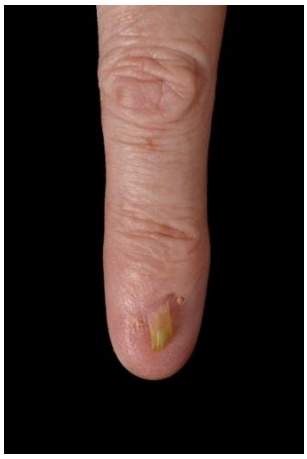
C



D



E



F

G

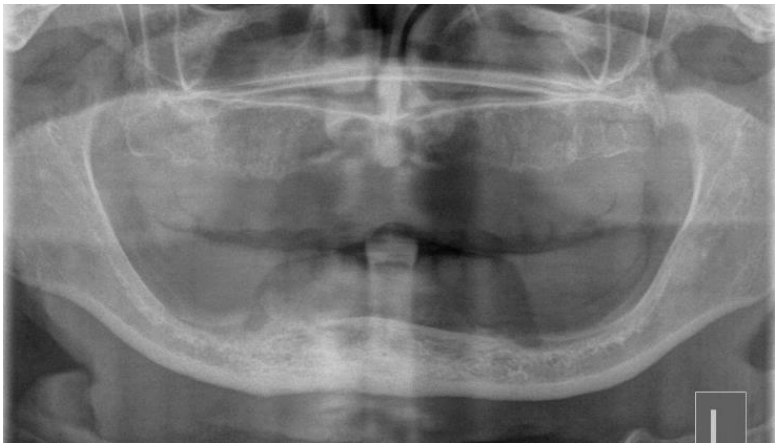


Figure 5.3: Phenotype of Halo68 (A-G) and her daughters (H-M). Halo68: note sparse eyebrows and thin hair (A-D), malformation of finger nail (E, F), and absence of teeth on an orthopantogram (G). One daughter has sparse eyebrows (H, I) but her hair (J), nails and teeth are normal. The second daughter is phenotypically normal (K-M). N and O: Halo47 wearing false teeth (N) and Individual 1.2 (O): both appear to have scanty eyebrow hair

5.4.3 *AXIN2:* c.1642G>T:p.Glu548*

The germline mutation identified in the affected siblings Halo47 and Halo68 is *AXIN2* c.1642G>T:p.Glu548* in exon 5. The mutation was confirmed to be present in the cDNA of Halo47 (4.1.4.2.3). If a truncated protein results from this mutation it is predicted to lose its DIX dimerization domain (http://www.uniprot.org/uniprot/Q9Y2T1#showFeaturesViewer accessed 04/05/2017).

5.4.4 Studies to Investigate the Mutation

The further studies performed to determine the pathogenicity of *AXIN2* c.1642G>T:p.Glu548* are described below.

5.4.4.1 Family Segregation Studies

There is a family history of colorectal neoplasia (Figure 5.2). It is important to ascertain whether the mutation is segregating with the disease phenotype. It is also important to try to determine which of the additional clinical features observed in the family, i.e. sparse eyebrows/ thin hair/ nail abnormalities/ oligodontia, are a part of the disease spectrum.

5.4.4.2 Protein Analysis to Confirm that the *AXIN2* Mutation Produces a Truncated Protein

It has already been shown that the mutation is present in DNA and is transcribed to RNA. This is interesting, as the location of the mutation is predicted to trigger NMD. NMD refers to the process in which mRNA harbouring premature termination codons (PTCs) is destroyed (Wen and Brogna 2008). It is a surveillance mechanism which prevents the accumulation of aberrant mRNA, and therefore potentially toxic truncated peptides (Wen and Brogna 2008). *AXIN2* c.1642G>T is predicted to result in NMD (http://www.mutationtaster.org/cgi-bin/MutationTaster/MutationTaster69.cgi accessed 31/08/2016) and yet it remains present in RNA, similar to the findings of Marvin *et al* (2011) for the reportedly

pathogenic mutation p.Tyr663*. The next logical step is to confirm that the mutation produces a truncated protein.

5.4.4.3 β-catenin IHC on FFPE Tumours to Assess Somatic Activation of the Wnt Pathway

APC mutations allow the accumulation and nuclear translocation of β-catenin. Histologically normal colorectal epithelia exhibit a membranous location of β-catenin, as observed with IHC (Wong *et al* 2004; Iwamoto *et al* 2000; Hao *et al* 1997; Valizadeh *et al* 1997). Colorectal adenomas, which are typically characterised by *APC* mutations, may display membranous and nuclear staining (Wong *et al* 2004; Iwamoto *et al* 2000; Hao *et al* 1997; Valizadeh *et al* 1997), but nuclear staining is not always observed (Kobayashi *et al* 2000). It is likely that nuclear staining increases with increasing degrees of dysplasia (Wong *et al* 2004; Iwamoto *et al* 2000; Kobayashi *et al* 2000; Hao *et al* 1997; Valizadeh *et al* 1997). β-catenin nuclear staining seems to occur in the majority of CRC (Wong *et al* 2004; Iwamoto *et al* 2000; Hao *et al* 1997), although not all papers report the same finding: Kobayashi *et al* (2000) observed that 55% of sporadic CRC exhibited negative/ scattered nuclear β-catenin expression. Different results could be attributed to differences in the protocols used for immunostaining, and differences in histological interpretation of neoplastic lesions (Wong *et al* 2004).

AXIN2 is a further component of the β-catenin 'destruction complex'. I therefore hypothesised that pathogenic *AXIN2* mutations might result in a similar pattern of β-catenin immunostaining as is observed in adenomas/ CRC with *APC* mutations. I sought to determine the subcellular location of β-catenin in adenomas/ CRC from patients with inherited variants in *AXIN2* using IHC. The results from this work are to be interpreted alongside the results from the somatic *APC* gene mutation screening in the same tumours.

5.4.4.4 Somatic *APC* Screening

The *APC* mutations which play a role in the pathogenesis of colorectal neoplasia are not random: 'typical' patterns are identified. Approximately a third of germline mutations in FAP patients occur at codons 1061 and 1309, and the reminder occur

largely uniformly between codons 1200 and 1600 (Fearnhead *et al* 2001). As described in the introduction chapter (1.5.1.1) of this thesis, in FAP the nature of the first mutation determines the type of second hit to *APC.* If the germline mutation occurs between codons 1194 and 1392, there is selection for allelic loss of *APC* as the second hit, whereas if the germline mutation lies outside of this region, the second hit is most likely to be a truncating mutation in the MCR (Fearnhead *et al* 2001). In sporadic tumours, over 60% of somatic mutations occur between codons 1286 and 1513, the MCR, which accounts for <10% of the coding sequence of *APC* (Fearnhead *et al* 2001). Within the MCR, there are further hotspots for mutations at codons 1309 and 1450 (reviewed in Fearnhead *et al* 2001) and codon 1554 (Rowan *et al* 2000).

APC mutations partially mediate their effects through enhanced Wnt-signalling. *AXIN2* mutations are hypothesised to have a similar effect. If this is true, then the tumours which occur in patients with inherited *AXIN2* mutations might not need to acquire *APC* mutations, so the 'typical' *APC* mutations observed in colorectal tumourigenesis might be absent. However, if typical *APC* mutations are identified in patients with germline *AXIN2* variants this could suggest that *AXIN2* variants are not pathogenic or alternatively that they are exerting tumourigenic effects through a mechanism other than through the Wnt-pathway.

To further elucidate the mechanism(s) by which *AXIN2* mutations may exert pathological effects, DNA was extracted from FFPE tumours originating in patients with inherited *AXIN2* variants. DNA was screened to determine whether 'typical' *APC* mutations were present.

5.4.4.5 *AXIN2* Loss of Heterozygosity (LOH) Analysis

LOH describes the situation whereby one allele of a gene is 'lost' somatically. It is a common genetic event in many cancers. It may develop through copy neutral mechanisms, in which there is a homologous recombination event or because a retained chromosome was duplicated, or it may be observed with copy loss, in which all/ part of the chromosome is lost (Ryland *et al* 2015). LOH is strongly associated with the loss of the wild-type allele in the tumours of individuals with inherited cancer predisposition syndromes who carry a germline mutation in a tumour suppressor gene (Ryland *et al* 2015).

Of the three papers which suggest that germline *AXIN2* mutations may predispose an individual to the development of polyposis +/- CRC (Lammi *et al* 2004; Marvin *et al* 2011; Rivera *et al* 2014), only one sought to investigate *AXIN2* LOH (Rivera *et al* 2014). Paradoxically it was the mutant allele which was lost in the tumours which were sampled. Two adenomas from the proband and four polyps (two adenomas, one HPP and one mixed polyp) from a sister were examined: LOH of the mutant allele was identified in one polyp from the proband, and in an adenoma and the mixed polyp from the sister. The authors also noted that the 'other adenoma (with low-grade dysplasia) had reduced amplification of one allele, and therefore incipient LOH, but not sufficiently so to confirm LOH *per se*'.

If *AXIN2* is acting as a typical tumour suppressor gene, and inherited mutations result in an increased risk of neoplasia, it would be expected that there might be loss of the wild type allele in tumours. This was investigated in DNA extracted from FFPE tumours.

5.4.4.6 Functional Characterisation of *AXIN2* c.1642G>T:p.Glu548*

Pathogenic mutations in *AXIN2* are proposed to result in increased activation of the Wnt-pathway, in a similar manner to *APC* mutations. To determine whether the variant identified in this study impacts Wnt-signalling, HEK293 TCF-Luc reporter cells were transfected with WT-*AXIN2,* the *AXIN2* mutation identified by Marvin *et al* (2011) (c.1989G>A, p.Tyr663* referred to hereafter as Fearon-*AXIN2.* See 5.3.2.7.1) or with the variant identified during this study (c.1642G>T:p.Glu548*, Short-*AXIN2)*. Wnt-pathway activation was assessed using luciferase assays.

5.4.5 Methods

5.4.5.1 Family Segregation Studies

In addition to Halo47 and Halo68, a further 6 family members were recruited to the study. The details of these individuals are in Figure 5.4 and Table 5.5 below:

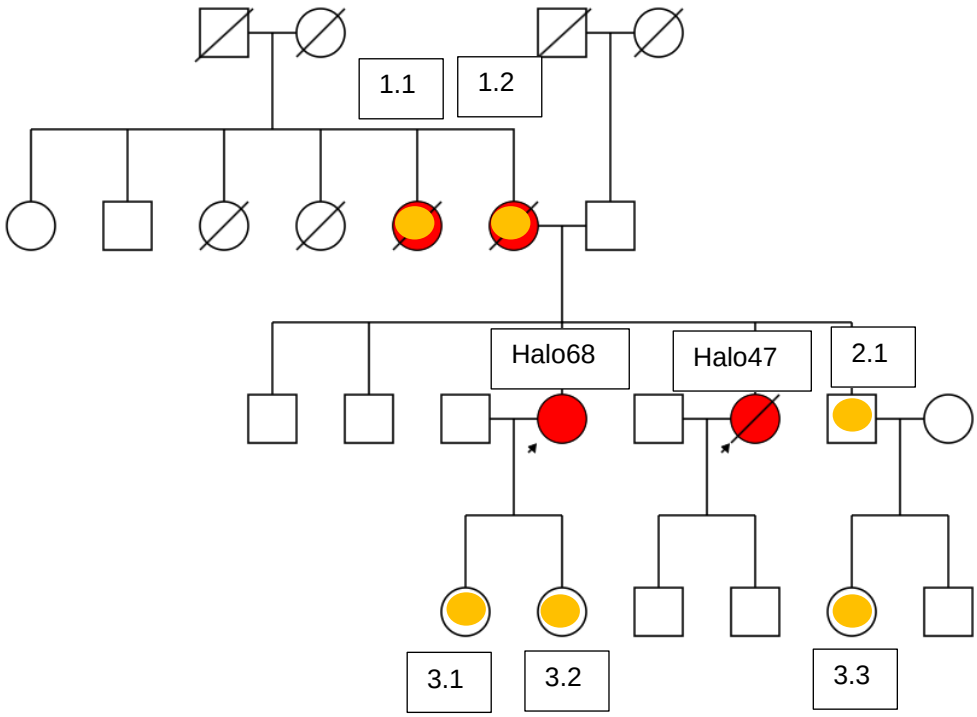


Figure 5.4: Family tree of Family A illustrating recruited individuals. The yellow circles indicate recruited family members.

Family Member	**Demographic Details**	**Phenotype**	**Material Available for Analysis**
Individual 1.1	Deceased female	Moderately differentiated CRC pT3N2Mx. 8 polyps identified macroscopically: 5 sampled: TA with LGD. Hemicolectomy performed when the patient was 79-years-old	FFPE material: tumour and normal mucosa (4 x TA LGD, 1 x CRC)
Individual 1.2	Deceased female	Moderately differentiated Duke's A CRC, a VA with LGD and 'three small metaplastic polyps', one of which was confirmed microscopically to be a HPP.	FFPE material: tumour and normal mucosa (1 x TA LGD, 1 x VA LGD, 1 x VA LGD that

		CRC diagnosed when the patient was 59-years-old. Allegedly did not develop adult teeth	CRC has arisen in, 1 x CRC, 1 x HPP)
Individual 2.1	48-year-old male	Clear colonoscopy in 2016 Normal teeth	DNA from whole blood
Individual 3.1	35-year-old female	Clear colonoscopy as a teenager, no recent colorectal examination Normal teeth	DNA from whole blood
Individual 3.2	27-year-old female	Never had a colonoscopy Normal teeth	DNA from whole blood
Individual 3.3	24-year-old female	Never had a colonoscopy Allegedly lacks 2 adult teeth	DNA from whole blood

Table 5.5: Demographic details and clinical phenotypes of recruited family members

For the individuals in whom blood was available, DNA was extracted from whole blood by the AWMGS (2.3.2). DNA was screened for the familial *AXIN2* mutation with PCR and Sanger sequencing using standard reagents and reaction conditions (2.3.8.1, 2.3.8.5, 2.3.9-2.3.14). The primers were those used for Haloplex validation (Appendix 4.2).

For the individuals in whom FFPE tissue was available, DNA was extracted from normal mucosa by the AWMGS using the Maxwell Promega LEV FFPE kit. DNA underwent PCR using reagents as per the protocol for PCR for DNA extracted from FFPE tissue (2.3.8.2, 2.3.8.5, 2.3.10). The primers were those used for Haloplex validation (Appendix 4.2), and the products were sequenced from reactions using Tas of 55.1 and 63.4^{0}C, which were the Tas resulting in successful PCR amplification. 6.25 µl of PCR product was used in the Big Dye reaction, which

followed standard conditions (2.3.11). Subsequent steps in the sequencing protocol followed the standard conditions (2.3.12, 2.3.13).

For any individual not found to carry the *AXIN2* mutation, the PCR and sequencing reactions were repeated with a second set of primers (Appendix 5.11).

5.4.5.2 Protein Analysis: Confirmation of a Truncated Protein

The familial *AXIN2* mutation, c.1642G>T:p.Glu548*, had been shown to be present in cDNA. To confirm that it produces a truncated protein, HEK293 TCF-Luc cells were transfected with *WT-AXIN2* plasmid, Short-*AXIN2* plasmid or empty plasmid. The Short-*AXIN2* plasmid had been generated through SDM of *WT-AXIN2* (5.3.2.6.1, 5.3.2.6.2). The same protocols and plasmid dilutions were followed as for the Luciferase assays (5.3.2.7, 2.5.7-2.5.10). A 24-well plate was used so the reaction volumes were adjusted accordingly. 48-hours after transfection, the cells were lysed and protein extracted (2.5.13). Protein was quantified using a BCA assay (2.5.14), and equivalent amounts of protein underwent Western Blotting, using anti-myc primary antibodies, and anti-mouse secondary antibodies (2.5.15). To ensure that equal amounts of protein had been loaded onto the electrophoresis gel, the membrane was stripped using Restore Plus Stripping Buffer (2.5.16) and was re-probed using primary antibodies against β-actin and anti-mouse secondary antibodies (2.5.15).

5.4.5.3 β-Catenin Immunohistochemistry (IHC)

FFPE tumours were available for Halo47 (CRC, TA LGD x 3, SA x 2) and Halo68 (TA LGD x 14, HPP x 6), their mother (Individual 1.2. CRC, the VA LGD that the CRC had arisen from, TA LGD x 1, VA LGD x 1, HPP x 1) and their aunt (Individual 1.1. CRC, TA LGD x 4).

IHC was also performed on a known FAP patient (germline mutation *APC* c.2940_2941 delA) as a positive control (CRC, TA LGD x 8, TA HGD x 1, mixed TVA/ HPP x1, biopsy of HGD/ CRC).

The Cellular Pathology Department, Cardiff and Vale NHS Trust, carried out β-catenin IHC using the EnVision FLEX System (Agilent) following the manufacturer's protocol and primary antibody Beta-Catenin Flex RTU.

5.4.5.4 Somatic *APC* Mutations in Colorectal Neoplasms

FFPE tumours were available for Halo47, Halo68, Individual 1.1 and Individual 1.2 as described in 5.4.5.3. DNA was extracted by the AWMGS using the Maxwell Promega LEV FFPE kit.

Due to a limited supply of poor quality DNA, *APC* MCR screening was performed, rather than the whole of the coding sequence. Codons 1181-1648 were sequenced. Work was carried out by the AWMGS following the protocol described in Appendix 5.12.

5.4.5.5 *AXIN2* Loss of heterozygosity Analysis (LOH)

FFPE tumours were available for Halo47, Halo68, Individual 1.1 and Individual 1.2 as described in 5.4.5.3. DNA was extracted by the AWMGS using the Maxwell Promega LEV FFPE kit.

LOH analysis was carried out by testing for heterozygosity at the site of the germline mutation. Standard reagents and reaction conditions were employed (2.3.8.2, 2.3.8.5, 2.3.9-2.3.13). The primers used are listed in Appendices 4.2 and 5.11. LOH analysis was performed for all the DNA samples which remained after somatic *APC* screening: several of the samples had been used to exhaustion.

5.4.5.6 Functional Characterisation of the *AXIN2* Variant

5.4.5.6.1 *AXIN2*-Containing Plasmids

Plasmids containing the human *AXIN2* coding sequence were very kindly supplied by Professor Eric Fearon, University of Michigan Medical School. The plasmid used during this study was N-terminal 6x*myc*-tagged *AXIN2* in pCS2+MT. Both wild-type *AXIN2* was provided, and mutant *AXIN2* containing the c.1989G>A variant described by Marvin *et al* (Marvin *et al* 2011), Fearon-*AXIN2*.

Plasmids were retrieved from filter paper (2.5.1) and used to transform XL-1 Blue Competent cells (2.5.4). Plasmid DNA was extracted using the MiniPrep kit (Qiagen) (2.5.5.) and the identity of the *AXIN2* coding sequence was confirmed with sequencing (2.5.5).

5.4.5.6.2 Site Directed Mutagenesis (SDM)

Site directed mutagenesis and bacterial transformation were employed to generate plasmids containing the *AXIN2* mutation identified as part of this study, c.1642G>T:p.Glu548* (2.5.6), Short-*AXIN2*. Primers are described in Appendix 5.13. Plasmid DNA was extracted using the MiniPrep kit (Qiagen) (2.5.5).
To confirm the presence of the desired mutation, sequencing was carried out by Dundee Sequencing. Primers are described in Appendix 5.14.

5.4.5.6.3 The Luciferase Reporter Assay in HEK293 TCF-Luc Cells

HEK293 TCF-Luc cells were kindly supplied by Professor Jeremy Nathans, John Hopkins University. They were kept at -80˚C. Once defrosted (2.5.7) they were seeded onto a 96-well Assay plate (2.5.9) prior to transfection with varying amounts of WT-*AXIN2,* Fearon-*AXIN2,* Short-*AXIN2* or WT/ Short-*AXIN2* plasmids (2.5.10). Following transfection, cells were stimulated with exogenous Human Wnt3a and Human R-Spondin-1 (2.5.11). WIST was subsequently added, before measuring WIST and luciferase activity (2.5.12).

5.4.6 Results

5.4.6.1 Family Segregation Studies

Six further family members were screened for the familial *AXIN2* mutation, c.1642G>T:p.Glu548*. The mutation was identified in Individuals 1.1 and 1.2, who

had a history of colorectal neoplasia. It was not present in the family members who were unaffected, Individuals 2.1, 3.1, 3.2, 3.3 (sequencing traces in Appendix 5.15)

5.4.6.2 Protein Analysis

HEK293 TCF-Luc cells were transfected with a series of dilutions of *WT-AXIN2,* Short-*AXIN2* or empty plasmid. Protein was extracted and analysed using Western Blotting.

Results showed that Myc-tagged-WT-AXIN2 was detected in the cells transfected with *WT-AXIN2* plasmid, and a truncated AXIN2 protein, approximately 30 KDa smaller than the WT protein, was detected in the cells transfected with Short-*AXIN2* plasmid. Protein was only detectable at the highest two concentrations of plasmid used for transfection. No AXIN2 protein was detected in the cells transfected with empty vector.

β-actin protein was detected with approximately equal intensity in all samples of protein lysate, confirming that the same amounts of protein had been loaded in each well.

The mutant AXIN2 was detected at greater intensity than WT protein, suggesting that it may possibly be a more stable protein. The protein ran at a larger size than predicted, by approximately 30 KDa, but this is commonly observed with AXIN2 (Dr. Anika Offergeld, personal communication). Possible reasons include it interacting with other proteins or the way in which it was interacting with the gel. See Figure 5.5.

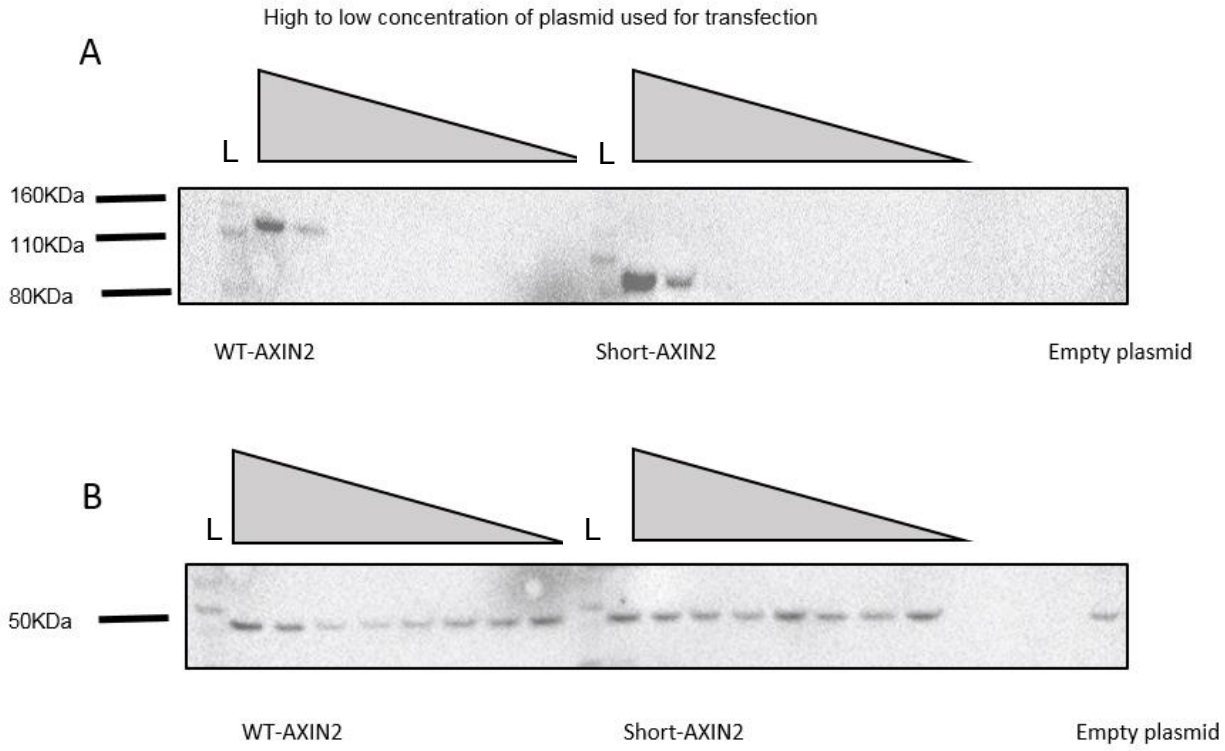


Figure 5.5. A: Western blot probing for AXIN2. L: ladder. Decreasing concentrations of AXIN2 plasmid used for cell transfection (1. 90 ng/25 µl, 2. 30 ng/25 µl, 3. 10 ng/25 µl, 4. 3.33 ng/25 µl, 5. 1.11 ng/25 µl, 6. 0.37 ng/25 µl, 7. 0.12 ng/25 µl, 8. ng/25 µl). B: Western blot probing for β –actin. L: ladder.

5.4.6.3 β-Catenin Immunohistochemistry (IHC)

Immunostaining was interpreted by myself and Professor Geraint Williams, and I am extremely grateful for his help and advice.

Results from the FAP positive control patient showed that β-catenin IHC in the adenomas exhibited largely membranous staining, but in most of the lesions, there were occasional cells with nuclear positivity. However, this did not seem to correlate with the degree of dysplasia, i.e. membranous staining was observed in both low-grade and high-grade regions, and the focal nuclear positivity present occurred in regions of low-grade dysplasia. However, LGD was by far the predominant morphology, so it is feasible that there wasn't sufficient HGD present to observe nuclear staining. The CRC exhibited mainly membranous staining of β-

catenin. There were occasional cells with nuclear positivity, and these tended to occur at the invasive margin of the tumour. See Figure 5.6.

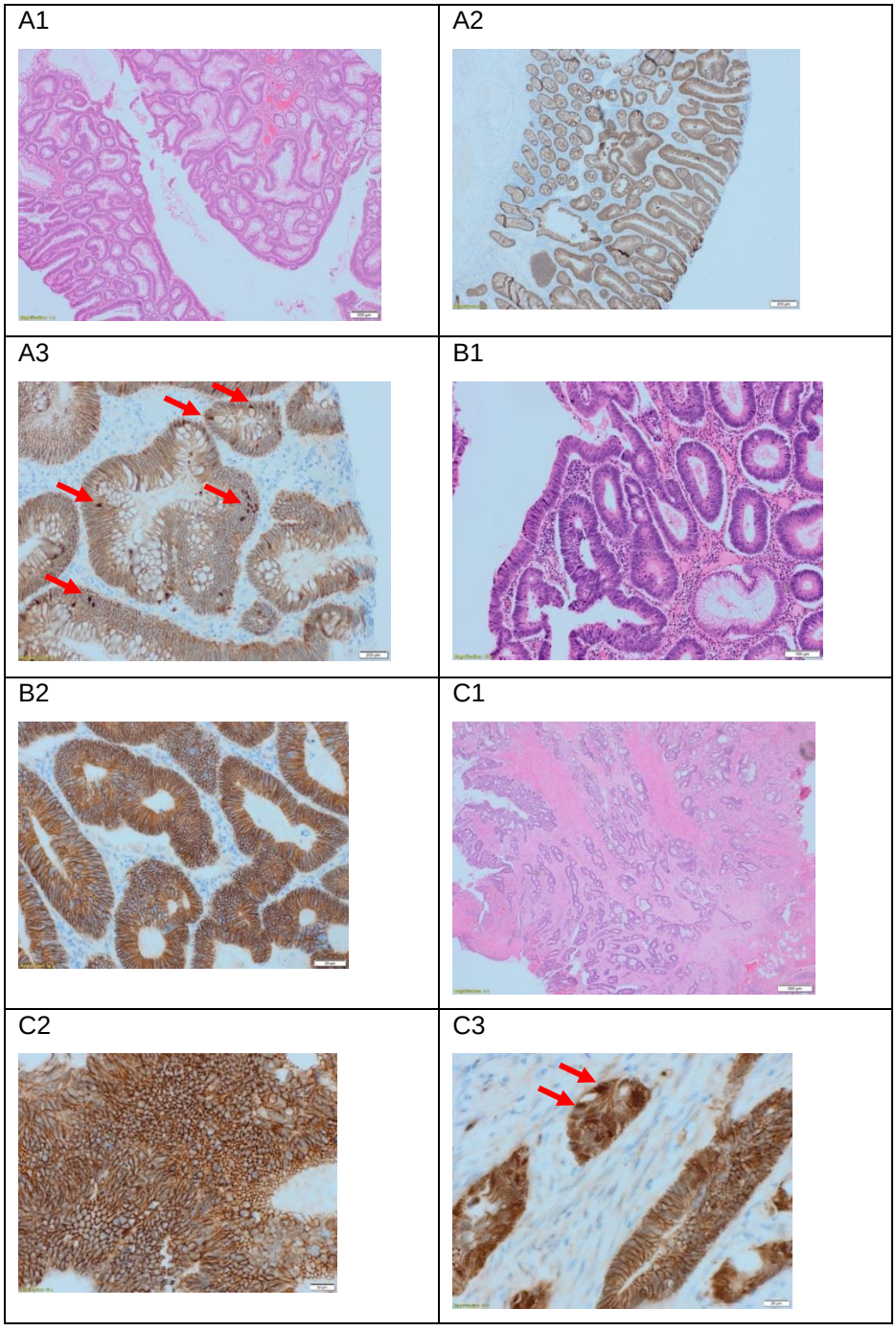


Figure 5.6: β-catenin IHC of colorectal adenomas and CRC from a patient with confirmed FAP, germline mutation APC c.2940_2941delA. A1: TA LGD H&E. A2: β-catenin IHC is largely membranous but there are scattered cells with nuclear staining as seen in A3 (red arrows). B1: Region of HGD H&E. Note membranous staining in B2. C1: CRC H&E. β-catenin IHC is predominantly membranous, C2, but there are occasional cells with nuclear positivity, C3 (red arrows). Scale bars in the bottom right corner of each image.

There was no evidence of nuclear β-catenin in any of the serrated/ hyperplastic lesions from patients carrying the *AXIN2* mutation (Figure 5.7):

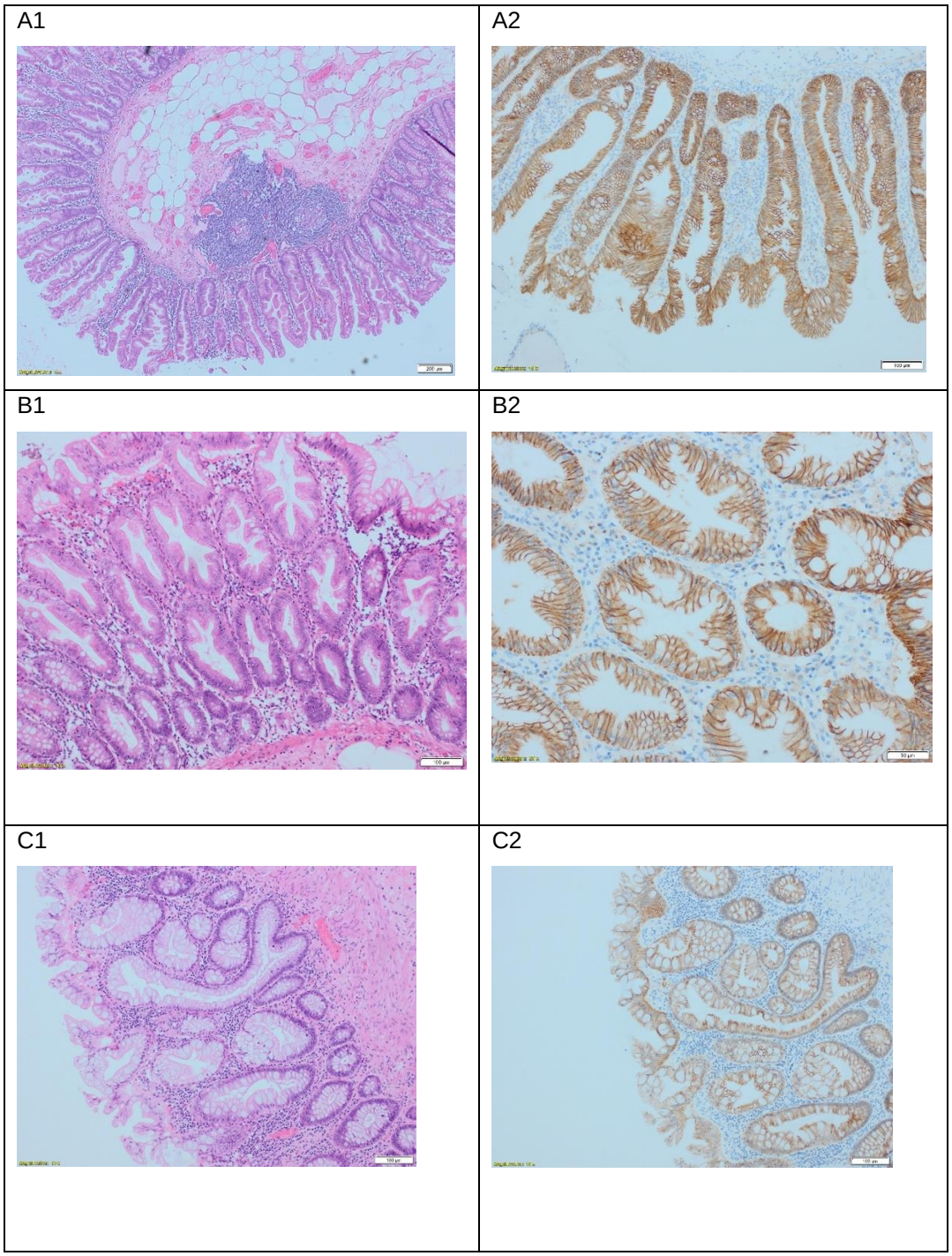


Figure 5.7: A1: Halo47 SA LGD H&E, A2. Halo47 SA LGD β-catenin IHC; B1: Halo68 HPP H&E, B2: Halo68 HPP β-catenin IHC; C1: Individual 1.2 HPP H&E, C2: Individual 1.2 HPP β-catenin IHC

All of the tubular adenomas from patients carrying the *AXIN2* mutation were low grade lesions. The prominent staining pattern was membranous. The only patient to exhibit any β-catenin nuclear positivity was Individual 1.1: in two adenomas there was an occasional cell with nuclear staining (Figure 5.8):

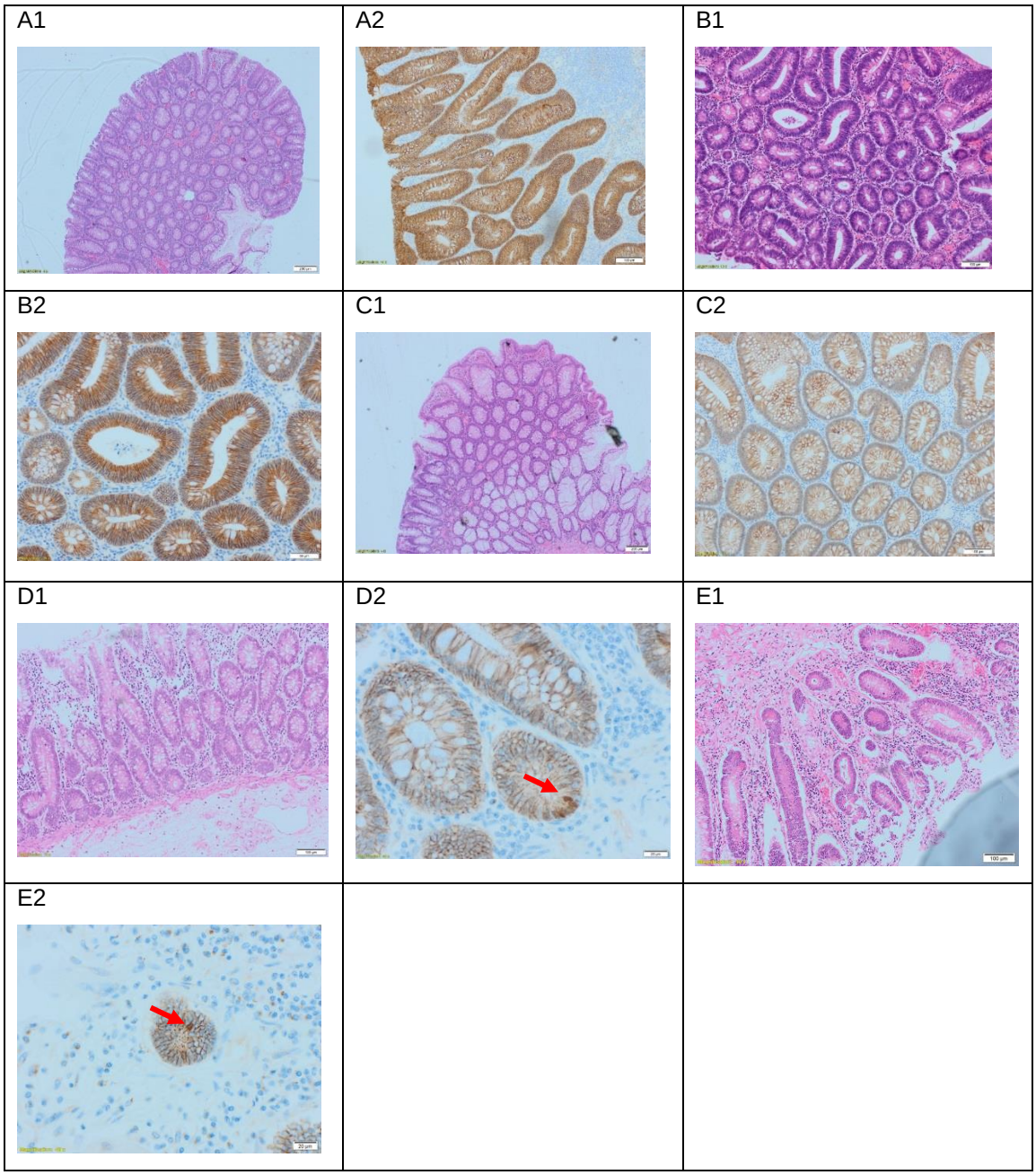


Figure 5.8: A1: Halo47 TA LGD H&E, A2 Halo47 TA LGD β-catenin IHC; B1: Halo68 TA LGD H&E, B2: Halo68 TA LGD β-catenin IHC; C1: 1.2 TA LGD H&E, C2: 1.2 TA LGD β-catenin IHC; D1: 1.1 TA LGD H&E, D2 TA LGD β-catenin IHC. Note focal nuclear staining (red arrow); E1: 1.1 TA LGD H&E, E2: TA LGD β-catenin IHC. Note focal nuclear staining (red arrow)

The only villous adenomata present were in Individual 1.2. Both the solitary VA and the VA which had progressed to carcinoma showed focal evidence of nuclear β-catenin staining. There was weak membranous staining of the carcinoma, but no evidence of nuclear β-catenin, although overall the staining pattern was very weak (Figure 5.9):

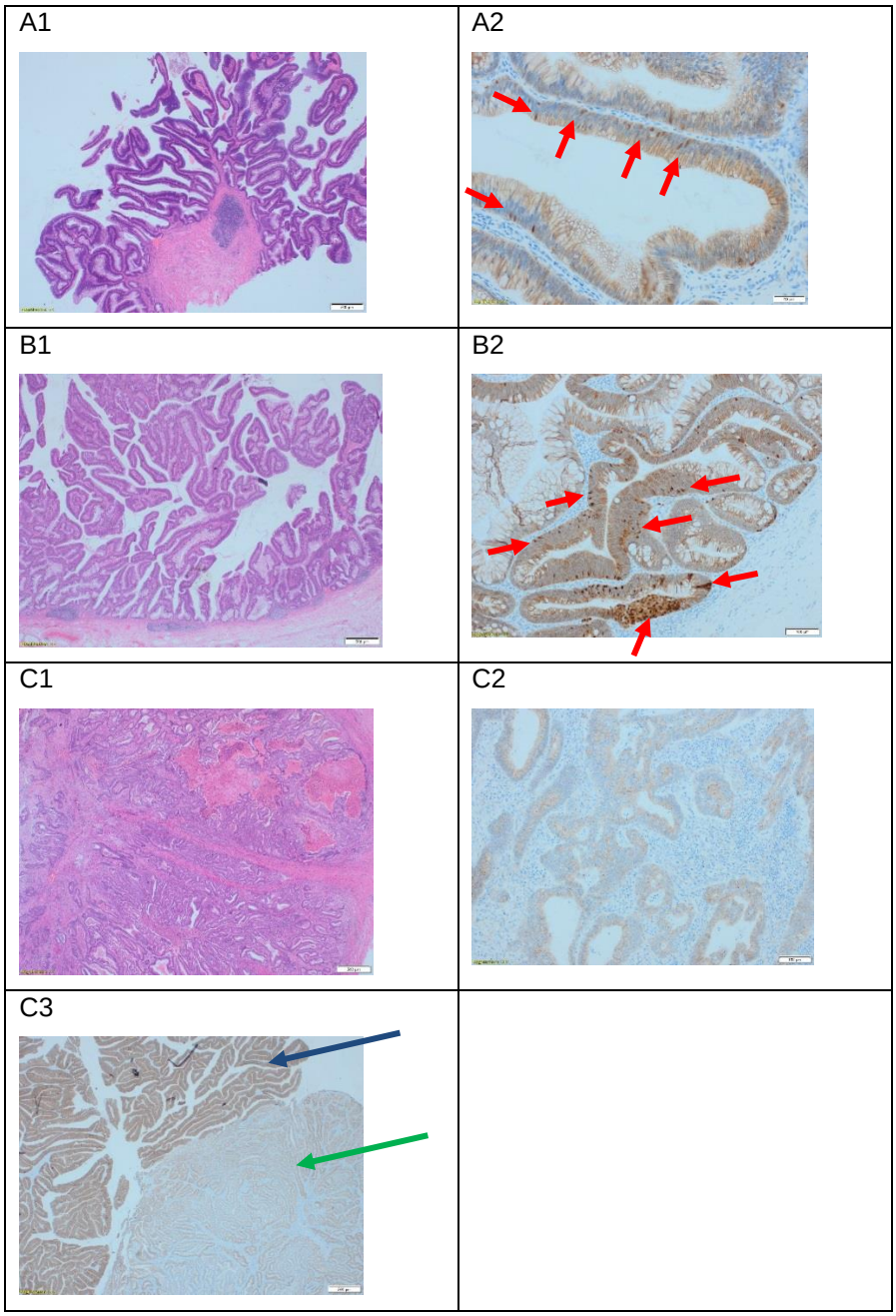


Figure 5.9: Individual 1.2 A1: Solitary VA H&E. A2: Solitary VA β-catenin IHC. Note focal nuclear staining (red arrows). B1: VA that has progressed to CRC H&E, B2: VA that has progressed to CRC β-catenin IHC. Note focal nuclear staining (red arrows). C1: CRC H&E. C2: CRC β-catenin IHC. C3: Contrast between the intensity of the staining of β-catenin between the villous adenoma (blue arrow) and the CRC (green arrow)

For Halo47, IHC staining of the adenocarcinoma showed different patterns: most of the tumour was moderately differentiated, but there were regions of poor differentiation. In the moderately differentiated regions, there were cells with prominent nuclear β-catenin staining, but this was not universal. In the poorly differentiated area, nuclear β-catenin was ubiquitous (Figure 5.10):

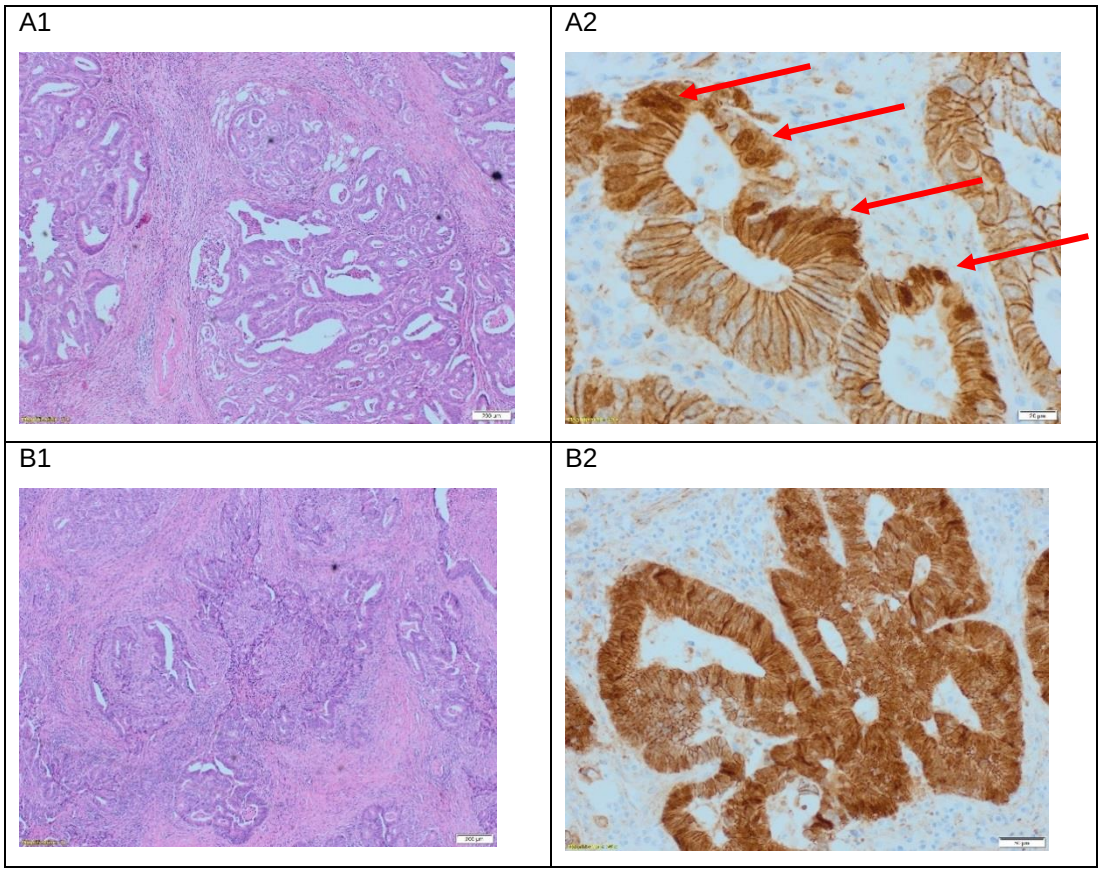


Figure 5.10: Halo47: A1: Moderately differentiated CRC H&E. A2: Moderately differentiated CRC β-catenin IHC. Note nuclear staining (red arrows) B1: Poorly differentiated CRC H&E. B2: Poorly differentiated CRC β-catenin IHC

The CRC in Individual 1.1 again showed predominantly membranous staining of β-catenin, but there were occasional cells with nuclear positivity (Figure 5.11).

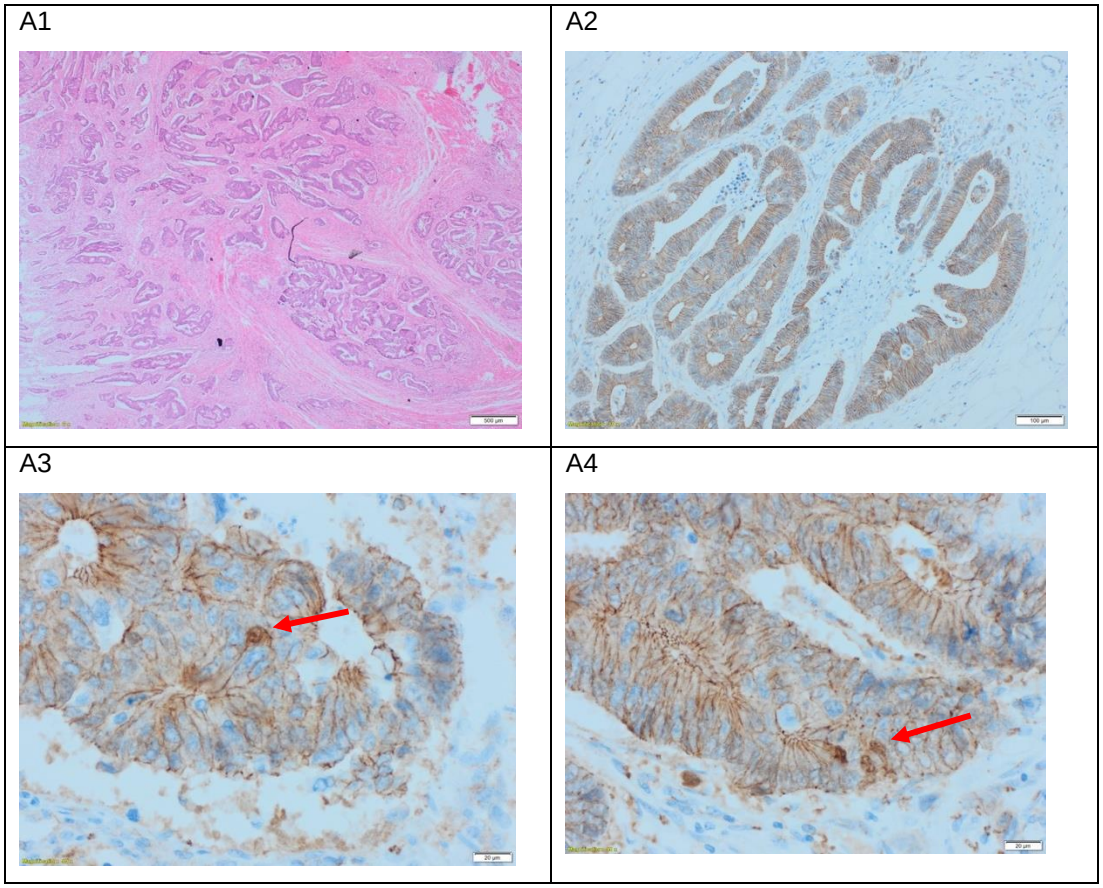


Figure 5.11: Individual 1.1: A1: Moderately differentiated CRC H&E; A2-A4: Moderately differentiated CRC β-catenin IHC. Focal nuclear staining is present (red arrows)

5.4.6.4 Somatic *APC* mutations in colorectal neoplasms

Several DNA samples extracted from FFPE tissue were of very low concentration, and PCR amplification followed by sequencing was not always successful. The successful PCR and sequencing reactions are described in Table 5.6. A pathogenic *APC* mutation refers to stop gains and frameshift mutations, it does not include silent or missense changes.

Sample ID	Case Reference	Diagnosis	Amplified PCR Fragments	Pathogenic *APC* Mutation
Halo47	H47-1	TA LGD	12/12	No
Halo47	H47-2	SA	9/12	No
Halo47	H47-3	SA	9/12	No
Halo47	H47-4	TA LGD	1/12	No
Halo47	H47-5	TA LGD	12/12	No
Halo47	H47-6	Mod. diff. CRC	11/12	No
Halo47	H47-7	Poorly diff. CRC	11/12	No
Halo68	H68-1	TA LGD	5/12	No
Halo68	H68-2	TA LGD	2/12	No
Halo68	H68-3	HPP	5/12	No
Halo68	H68-4	TA LGD	4/12	No
Halo68	H68-5	TA LGD	2/12	No
Halo68	H68-6	TA LGD	4/12	No
Halo68	H68-7	HPP	2/12	No
Halo68	H68-8	TA LGD	2/12	No
Halo68	H68-9	TA LGD	2/12	No
Halo68	H68-10	TA LGD	1/12	No
Halo68	H68-11	HPP	1/12	No
Halo68	H68-12	HPP	1/12	No

Halo68	H68-13	HPP	2/12	No
Halo68	H68-14	TA LGD	0/12	N/A
Halo68	H68-15	HPP	0/12	N/A
Halo68	H68-16	TA LGD	6/12	No
Halo68	H68-17	TA LGD	4/12	No
Halo68	H68-18	TA LGD	6/12	No
Halo68	H68-19	TA LGD	3/12	No
Halo68	H68-20	TA LGD	5/12	No
1.2	1.2-1	VA LGD	3/12	No
1.2	1.2-2	CRC, mod. diff.	9/12	No. However, the fragment containing p.Glu1286* in 1.2-3 failed to amplify
1.2	1.2-3	VA LGD (has progressed to CRC)	10/12	c.3856G>T: p.Glu1286*
1.2	1.2-4	HPP	10/12	No
1.2	1.2-5	TA LGD	7/12	No
1.1	1.1-1	TA LGD	0/12	N/A
1.1	1.1-2	TA LGD	7/12	No
1.1	1.1-3	TA LGD	8/12	c.3870_3877delTCAGACGA
1.1	1.1-4	CRC, mod. diff.	9/12	c.3667delT

Table 5.6: Somatic APC screening results

It was therefore seen that only a minority of tumours were identified as having a somatic *APC* mutation present.

5.4.6.4.1 Correlation between β-Catenin Immunohistochemistry (IHC) and Somatic *APC* Mutations

Only a small proportion of lesions exhibited evidence of nuclear β-catenin or were identified as having a somatic *APC* mutation. The correlation between these findings is described in Table 5.7.

Patient	**Sample Morphology**	**Nuclear β-Catenin**	**Somatic *APC* Mutation**
Halo47	CRC	Yes	Uncertain (11/12 fragments sequenced)
1.1	TA LGD	Yes	Uncertain (7/12 fragments sequenced)
1.1	TA LGD	Yes	Yes c.3870_3877delTCAGACGA
1.1	CRC	Yes	Yes c.3667delT
1.2	VA	Yes	Uncertain (3/12 fragments sequenced)
1.2	VA progressed to CRC	Yes	Yes c.3856 G>T

Table 5.7: Correlation between nuclear β-catenin IHC and the presence of a somatic APC mutation

5.4.6.5 *AXIN2* Loss of Heterozygosity Analysis (LOH)

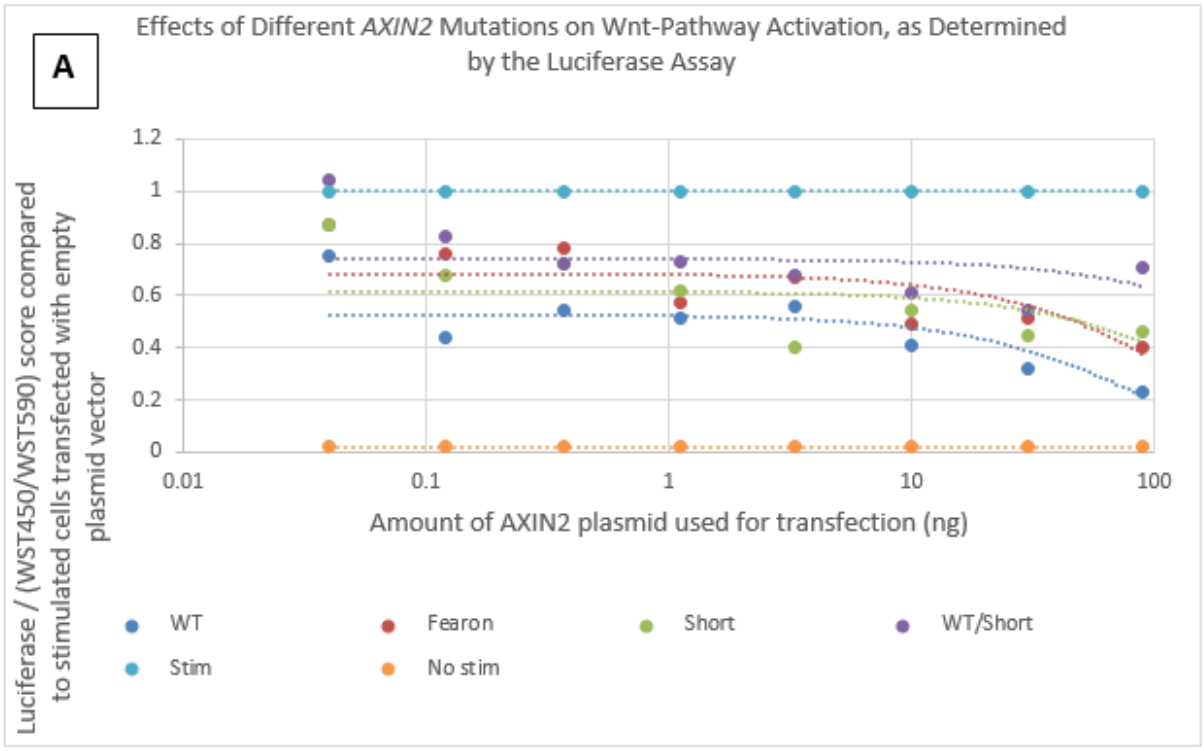
Successful LOH screening was achieved for a total of 12 tumours/ tumour regions. There was no evidence of LOH of the WT *AXIN2* allele in any of the tumours sampled (sequencing traces in Appendix 5.16).

5.4.6.6 Functional Characterisation of the *AXIN2* Variant

In order to further characterise the effects of the *AXIN2* c.1642G>T:p.Glu548* variant, studies were performed in HEK293 TCF-Luc cell lines. Luciferase assays were used to determine the effects of the mutation on Wnt-pathway activation.

Cultured cells transfected with empty plasmids were stimulated with exogenous Wnt3a and R-Spondin. This lead to an approximately 67-fold increase in the normalised luciferase score, compared to unstimulated cells.

Cells were transfected with a series of dilutions of *AXIN2* plasmid: WT-*AXIN2*, Fearon-*AXIN2,* Short-*AXIN2* or a WT/ Short mix. At all concentrations of plasmid, mutant *AXIN2*, both Fearon and Short, was less able to inhibit reporter gene activity than WT-*AXIN2* (Figure 5.12A). Interestingly, it was the WT/ Short mix which had the least effect on Wnt-pathway inhibition.



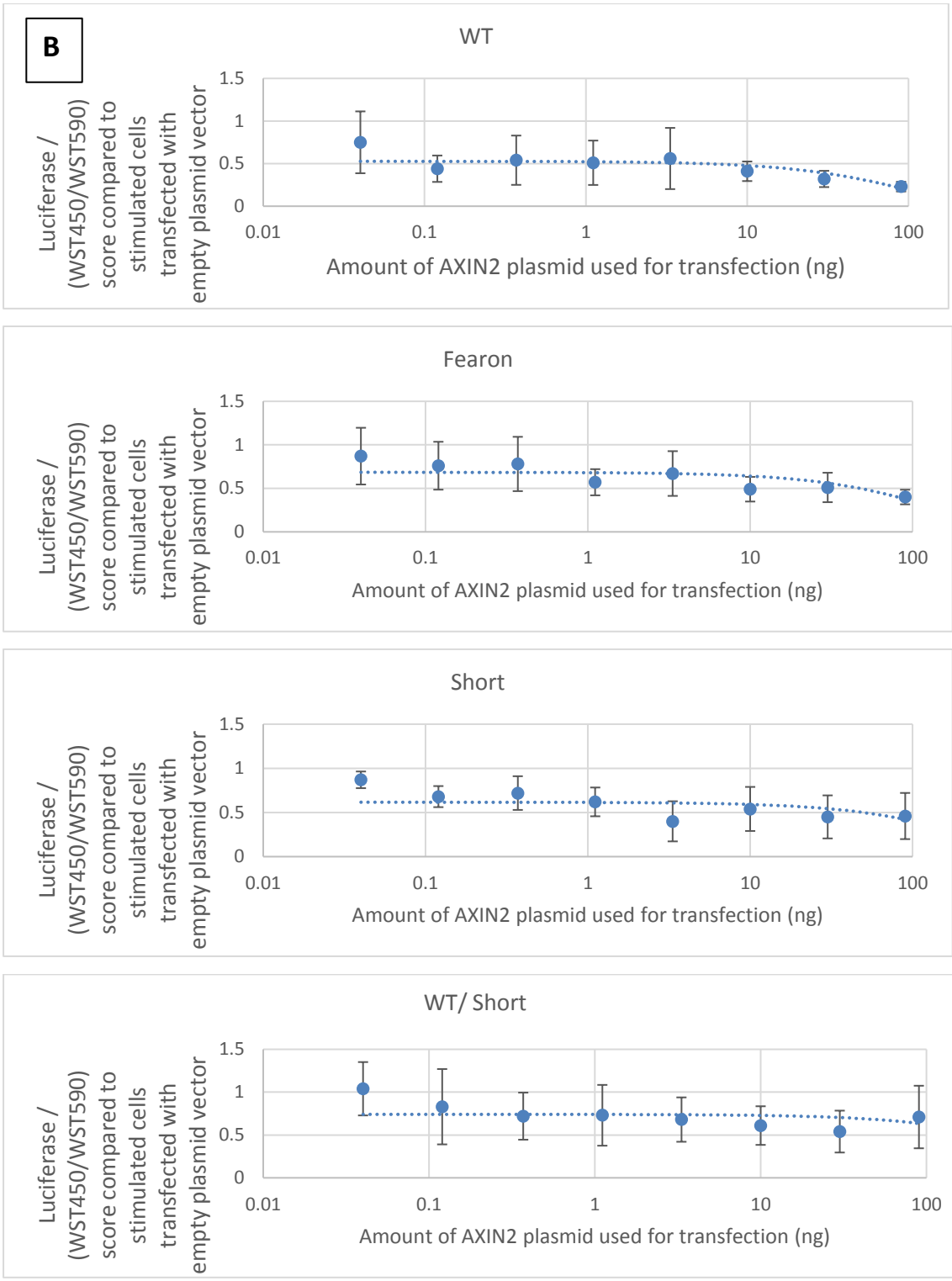


Figure 5.12A: Graph to show the different effects of AXIN2 mutations on Wnt-pathway inhibition. Mutant AXIN2 is less able to inhibit Wnt-pathway activation than WT-AXIN2; 5.12B: Standard deviations shown for each separate plasmid used for transfection.

Statistical analysis was carried out with the help of Dr. Matthew Mort. A 2-way ANOVA analysis and Tukey multiple pairwise analysis was performed to determine whether the difference in Wnt-pathway inhibition was statistically significant between the WT-*AXIN2* and the mutant plasmids. Statistical significance was reached when WT-*AXIN2* was compared to the Short/ WT- *AXIN2* mix, with a p-value of 0.00155. Statistical significance was not reached for Fearon-*AXIN2* (p=0.1021872) or Short-*AXIN2* (p=0.1274879) compared to WT-*AXIN2.* This could be due to the small sample size.

5.4.7 Conclusions of Studies into *AXIN2* c.1642G>T:p.Glu548*

The *AXIN2* c.1642G>T:p.Glu548* mutation was initially identified in the sibling pair Halo47 and Halo68, both of whom had been recruited to the study due to their phenotype of colorectal polyposis. The same variant was subsequently identified in their mother and maternal aunt. Both had colorectal neoplasia: they had a history of CRC, but their polyp counts were below 10. The variant was not present in a brother, who had recently had a clear colonoscopy, and in the daughters and niece of Halo68.

Both sisters and their mother allegedly lacked all adult teeth. This was confirmed by dental X-rays in Halo68. We were unable to get any information about the dental history of the aunt. Halo68 also exhibited clinical signs of ectodermal dysplasia – she had sparse hair on the outer half of her eyebrows, thin hair on her head and a nail malformation. We were unable to do a clinical assessment of the remaining variant-carriers in the family as they had all passed away, but family photos of Halo47 and her mother were suggestive of sparse eyebrow hair. Whether these features are part of the phenotypic spectrum associated with *AXIN2* mutations remains to be determined. In both Lammi's (Lammi *et al* 2014) and Marvin's (Marvin *et al* 2011) papers, *AXIN2* mutant-carriers had oligodontia, and in the family reported in Marvin's paper, signs of ectodermal dysplasia were also present. In this study, a daughter of Halo68 had sparse eyebrows and no other signs of ectodermal dysplasia, but did not carry the familial mutation. It seems that oligodontia segregates with the colorectal phenotype, but further studies in *AXIN2*-mutant families are required to better characterise the clinical features.

Western blotting confirmed that cell lines transfected with Short-*AXIN2* plasmids produce a protein approximately 30 KDa smaller than cells transfected with *WT-AXIN2* plasmids. The truncated protein may possibly be more stable than WT protein. This would need to be confirmed by a technique such as a pulse chase experiment. Increased stability of mutant AXIN2 had previously been reported for the truncated protein investigated by Mazzoni *et al* and may be due to the loss of a destabilising phosphorylation site (Mazzoni *et al* 2015).

It is possible that the mutation may have further effects in addition to being a nonsense change. Analysis using MutPred Splice (http://www.mutdb.org/mutpredsplice/about.htm accessed 06/03/2018) suggests that the mutation is a splice neutral variant with a score of 0.37. However, *in silico* analysis using Human Splicing Finder (http://www.umd.be/HSF3/ accessed 06/03/2018) suggests that the mutation could possibly add an exon splicing silencer and remove an exonic splicing enhancer. In theory, this could lead to skipping of Exon 5. Analysis with Mutalyzer (https://mutalyzer.nl accessed 06/03/2018) confirmed that the loss of Exon 5 would produce a truncated protein, but that the transcript would be in-frame. The truncated product would be approximately 88 KDa. Exon 5 is partially responsible for encoding the domain of AXIN2 which interacts with GSK3β (http://www.uniprot.org/uniprot/Q9Y2T1#showFeaturesViewer accessed 06/03/018) so if it was absent from the protein, this could have functional consequences. To further investigate this, techniques such as mass spectrometry could be used.

β-catenin IHC was initially performed on tumour samples from a known FAP patient. Throughout all adenomas and CRC, the staining pattern was predominantly membranous, with occasional cells exhibiting nuclear positivity. For the patients carrying the familial *AXIN2* mutation, Halo47, Halo68, Individual 1.1 and Individual 1.2, there was no nuclear staining in any of the hyperplastic lesions. This is as expected, as such lesions do not typically exhibit activation of the Wnt-pathway. Of all the tubular adenomas present, only two had any focal nuclear β-catenin staining. It was anticipated that if the adenomas had similar levels of Wnt-pathway activation as do *APC*-mutant tumours, there should have been more evidence of nuclear positivity. Interestingly, the villous adenomata, which are more likely to progress to cancer than tubular lesions did show evidence of nuclear β-catenin staining.

There were three CRC present – in Halo47, Individual 1.1 and Individual 1.2. The cancer in Halo47 showed different patterns of β-catenin staining: most of the tumour was moderately differentiated, but there were regions of poor differentiation. In the moderately differentiated regions, there were cells with prominent nuclear β-catenin, but this was not universal. In the poorly differentiated area, nuclear β-catenin was ubiquitous. In Individual 1.1, the cancer again showed predominantly membranous staining of β-catenin, but there were occasional cells with nuclear positivity. In Individual 1.2, the cancer had arisen in a VA. There was weak membranous staining of the carcinoma, but no evidence of nuclear β-catenin. However, the staining of the cancer was repeatedly very weak. The reason for this is not clear. One possibility is that β-catenin had acquired a mutation which rendered it relatively resistant to IHC staining. Regardless of the aetiology, the poor staining makes interpretation difficult.
Nuclear positivity for β-catenin in the VA and CRC, but not in most of the TA, implies that Wnt pathway activation may be occurring at a later stage of tumourigenesis than in FAP adenomas.

Somatic *APC* analysis was performed to determine whether 'typical' *APC* mutations were present in tumours from patients carrying germline *AXIN2* mutations. Although DNA was extracted from 36 lesions, less than half of the desired PCR fragments were successfully amplified, making interpretation difficult. Only 3 truncating mutations were identified: 1 of these was present in cancer and another was present in the VA which had developed into CRC. This raises the possibility that adenomas arising in *AXIN2* mutation carriers might need to acquire mutation(s) in a gene such as *APC* in order to acquire sufficient Wnt pathway activation to progress to malignancy.

We attempted to determine whether we had identified statistically significantly fewer *APC* mutations than would have been predicted in adenomas from *AXIN2* mutation carriers. However, it was not possible to perform such an analysis due to the lack of a robust control set to compare our results to. Whilst there is a wealth of data in the literature documenting somatic mutations in sporadic CRC and in FAP-associated adenomas and CRC, there are scanty papers addressing somatic changes in sporadic adenomas. Of the papers that do exist, different experimental approaches have been taken. Some studies considered the whole of the coding region of *APC*, some look at exon 15 only and others address specific nucleotide sequences; some

of the studies sequence fresh tissue, some use frozen tissue and others use FFPE material; some studies investigate sporadic lesions only, others include some FAP-associated tumours (Van Wyk *et al* 1999; Olschwang *et al* 1998; De Benedetti *et al* 1994; Miyoshi *et al* 1992b; Powell *et al* 1992).

The correlation between a somatic *APC* mutation with the presence of nuclear β-catenin staining was addressed: in both of the cancers and the TA LGD which had identified somatic *APC* mutations (Individual 1.2 and Individual 1.1) nuclear staining was present. There were tumours with nuclear β-catenin staining which did not seem to have a somatic *APC* mutation. However, in all of these samples, the MCR was not successfully completely sequenced, so the actual somatic *APC* MCR mutation status is uncertain.

It appears that tumours arising in individuals carrying the germline *AXIN2* c.1642G>T:p.Glu548* mutation may not exhibit the 'typical' *APC* mutations seen in patients with FAP or sporadic adenomas, although the poor coverage of the MCR makes it difficult to say this with certainty. If this is true, it implies that lesions follow a different genetic tumourigenic pathway. This exciting possibility is going to be further investigated by carrying out WES on tumour DNA, but this was not possible during the timescale of this project.

There was no evidence of LOH in any of the tumours sampled. If *AXIN2* was a typical tumour suppressor gene, and if the variants identified are pathogenic, it would be anticipated that there would be loss of the wild-type allele in the tumours. This was not observed. It is possible that the mutation is exhibiting dominant negative effects. It is also possible that there is a second hit to the WT allele by a mechanism other than LOH, for example epigenetic silencing or a second mutational event.

The effects of the Fearon and Short *AXIN2* mutations on Wnt-pathway activation were examined through the use of luciferase assays in HEK293 TCF-Luc cell lines. Fearon-*AXIN2* and Short-*AXIN2* both demonstrated possibly impaired Wnt pathway inhibition, but this did not reach statistical significance. Interestingly, the plasmids least able to inhibit the Wnt-pathway were the Short/ WT mix. This would be consistent with the notion that inherited *AXIN2* mutations can act via a dominant negative mechanism. Previous papers have shown that mutant AXIN2 is still able

to dimerise with itself and with WT-AXIN2 (Mazzoni *et al* 2015). If the same is true for the mutation we have identified, it is possible that a complex composed of WT and mutant AXIN2 is less efficient at promoting β-catenin degradation than mutant AXIN2 alone.

Halo47 and Halo68 have also undergone WES of blood DNA (4.3). This did not identify any inherited variant that was considered more likely to be clinically significant than the confirmed *AXIN2* mutation. In particular no likely pathogenic variants were identified in known adenomatous polyposis or ectodermal dysplasia genes.

Overall the results from the studies performed would be consistent with some inherited *AXIN2* mutations having a role in the pathogenesis of a colorectal polyposis syndrome in which individuals also display signs of ectodermal dysplasia, in particular oligodontia. If confirmed, this apparently very rare syndrome could be named AxAP: *AXIN2*-Associated Polyposis. The effects of mutant *AXIN2* are likely to be at least partially mediated through activation of the Wnt-signalling pathway. It is important that further studies are performed to confirm and characterise the syndrome, so that optimal clinical management can be determined. Such further studies would include an improved genetic characterisation of the tumours arising in patients with inherited *AXIN2* mutations to better understand the molecular pathways of tumourigenesis in these individuals. Currently the WGP is developing a platform to allow WES on DNA extracted from FFPE tissue, which will be performed later this year. It is also important to gain a greater knowledge about any other effects which mutant *AXIN2* may have, for example on the numbers of stem cells in tissue, its effects on cell proliferation and its effects on other signalling pathways.

The *AXIN2* mutation identified in this study, c.1642G>T:p.Glu548*, is not present in Exac, HGMD or in data from the 1000 Genomes Project. However, somatic *AXIN2* mutations are well described in a variety of cancer types. The COSMIC database reports 22 somatic nonsense mutations of *AXIN2* in 46376 cancer samples tested: 10 of these were large intestine adenocarcinoma (https://cancer.sanger.ac.uk/cosmic/gene/samples?all_data=&coords=AA%3AAA&dr=&end=844&gd=&id=4988&ln=AXIN2&mut=substitution_nonsense&seqlen=844&src=gene&start=1 accessed 15/05/2018). CBioPortal reports that 1.4% of cancers screened had a somatic mutation of *AXIN2.* 243 truncating mutations are

described: the vast majority of these were cancers in the gastrointestinal tract (http://www.cbioportal.org/index.do?session_id=5aec70b3498eb8b3d565fab9 accessed 15/05/2018). In the Exac database, there are 9 entries of loss of function germline mutations in *AXIN2.* They are all rare – the commonest is a frameshift, p.Asn666Glnfs*41, which has an allele frequency of 0.0005561 and was identified in 65/116894 alleles, being overrepresented in African populations. Although the phenotype of these patients is not known, Lammi's paper (2004) describes a frameshift mutation in a 13-year old boy with oligodontia, which truncates AXIN2 at almost the same location: the group reported a heterozygous 1 bp insertion after nucleotide 1994 in exon 7, which expands one of the several mononucleotide repeats. It recodes the amino acids starting at Asn666 and incorporates a stop codon 40 codons later. On CBioportal there are 6 reports of a stop codon being inserted 41 codons after p.Asn666 in CRC (http://www.cbioportal.org/index.do?session_id=5aec70b3498eb8b3d565fab9 accessed 15/05/2018). The significance of the frameshift reported in Exac is therefore not determined. It is possible that it is associated with a phenotype that we are not aware of, or alternatively it may be well-tolerated, possibly because it occurs relatively close to the C terminus of the protein.

It is interesting that the colorectal phenotype in patients carrying reported specific *AXIN2* mutations seems to be very variable, and both adenomas and HPPs seem to form part of the clinical spectrum. The consistent observation in this and previous reports is that patients with polyposis and truncating *AXIN2* mutations also have oligodontia. Oral abnormalities are common in FAP patients, with a prevalence of 17% (reviewed in Groen *et al* 2008). Such abnormalities include supernumerary teeth, dentigerous cysts and secondary retention of teeth (reviewed in Groen *et al* 2008), likely due to aberrations in Wnt signalling. The canonical Wnt pathway plays an essential role in tooth initiation and morphogenesis, and also in dental cell differentiation (reviewed in Järvinen *et al* 2018). In *AXIN2* mutant patients, individuals appear to grow their primary dentition (baby teeth) but do not develop a full complement of secondary teeth. Järvinen *et al* (2018) sought to investigate the effects of the Wnt pathway on the formation of molars in mouse models and in *ex vivo* cultures. They showed that increasing Wnt/ β-catenin activity in dental mesenchyme inhibited the development of posterior molars, whereas reducing activity was associated with continued tooth development (Järvinen *et al* 2018). The authors concluded that the dental abnormalities observed in *AXIN2* mutant

patients results from the modulation of Wnt/ β-catenin signalling in the dental mesenchyme (Järvinen *et al* 2018).

In mice studies of *AXIN2*, homozygous knockout of *AXIN2* leads to reduced head growth, due to premature fusion of cranial sutures. This resembles craniosyntosis in humans (Yu *et al* 1995). The dental and colorectal phenotype in such models is not described.

5.5 Chapter Conclusions

The aim of this chapter was to carry out genetic, *in silico* and functional assessment of mutations identified through targeted exome sequencing (Chapter 4). Variants were selected in the MMR genes, *POLE*, *POLD1* and *AXIN2.*

Tumours from patients carrying MMR gene variants underwent MMR IHC and MSI testing. There was no evidence of loss of MMR proteins or microsatellite instability in any of the lesions. This leads to the conclusion that the MMR gene variants identified are unlikely to be clinically significant.

Four patients were found to have germline variants in *POLE/ POLD1.* These were investigated through a combination of protein modelling, somatic 'hotspot' mutation screening and mutation signature analysis. One of the variants, POLE p.Glu2140Lys in Halo18, is reported as benign on dbSNP. It does not lie within any of the POLE functional domains and was not in the protein model used as part of this project. The variant is presumed not to be clinically significant. A POLE variant, p.Tyr623Cys in Halo61, remains a VUS. Although mutation signature analysis was attempted, this was not successful. Somatic mutation analysis is still in progress for tumours from Halo15 and Halo27. Although this should be complete in the near future, at the present time the mutations identified in Halo15 and Halo27, c.1411A>G:p.Met471Val and c.2861C>T:p.Thr954Met, remain as VUS.

A previously unreported truncating *AXIN2* mutation, c.1642G>T:p.Glu548*, was identified in a family with a 2-generation history of oligodontia and colorectal neoplasia. The mutation segregates with the disease phenotype in an AD manner. It is present in DNA and cDNA and produces a truncated protein. P.Glu548* mutant AXIN2 appears less able to inhibit the Wnt pathway than WT AXIN2 and could have dominant negative effects.

The study of Family A has added to evidence from other previously reported familes with *AXIN2* mutations that there may exist a rare *AXIN2*-Associated Polyposis Syndrome, in which individuals have signs of ectodermal dysplasia and a variable colorectal phenotype. The effects of the mutant protein could be at least partially mediated through activation of the Wnt signalling pathway. Further studies are required to confirm and better characterise this syndrome. Research programmes, including the UK 100 000 Genomes Project that is recruiting patients with colorectal polyposis, may identify further affected patients with germline *AXIN2* mutations.

Chapter 6 Thesis Discussion

Colorectal carcinoma (CRC) is the fourth most common cancer in the United Kingdom (UK) and the second most common cause of cancer death (Cancer Research UK 2015). The majority of colorectal carcinomas occur sporadically, but in 15-35% of patients, hereditary factors are important (reviewed in Mishra and Hall 2012; Burt 2007). These factors include the inherited polyposis syndromes, FAP, MAP, PPAP, *NTHL1*-Associated Polyposis and *MSH3*-Associated Polyposis. In all of these syndromes, affected individuals develop multiple colorectal polyps and have an increased risk of developing CRC. In order to assess whether an inherited disorder is present in patients with multiple polyps, it has been traditional practice to refer individuals with >10 polyps to a medical genetics service for genetic counselling and consideration of genetic testing. In the NHS, *APC* and *MUTYH* are routinely screened, with increasing numbers of laboratories also screening for the recurrent mutations in *POLE* and *POLD1.* Of the cohort of individuals who undergo genetic diagnostic testing, up to 50% do not have a mutation identified. It is this group of patients that this thesis has addressed.

The key hypotheses thought to underlie the patients' phenotypes were that mutations may be present outside of the ORF of *APC/ MUTYH*, that the affected individuals may exhibit somatic mosaicism of *APC* or that further polyposis genes may play a role. This thesis has not considered epigenetic effects or oligogenic inheritance.

To address these hypotheses, firstly *APC* and *MUTYH* were interrogated. UDS of the entire genomic loci was performed, qPCR analysis was utilised to study gene transcription, and patient-derived RNA samples underwent sequencing and AI analysis.
The next stage was to carry out targeted exome sequencing of 15 genes either known to be, or possibly, involved in the pathogenesis of colorectal neoplasia. Following on from this, a subset of patients underwent WES. The final phase of the project was to carry out further genetic, *in silico* and functional studies on the candidate variants most likely to be clinically significant, to obtain further evidence to support or refute their role in colorectal polyposis.

6.1 *APC* and *MUTYH* Studies

The most interesting results from the examination of *APC* and *MUTYH* were those from the gene expression studies. Four patients appeared to have evidence of reduced *APC* transcription. These individuals all had phenotypes in keeping with a clinical diagnosis of FAP. Their gene transcription results were consistent with their AI results, which either showed apparent loss of transcription of one *APC* allele (Halo52 and Halo53) or were uninformative (Halo46 and Halo64). The cause of the reduced *APC* expression was determined in Halo46. UDS identified a 5'UTR variant (G>A at 5:112043225), at position -c190. Subsequent testing of family members confirmed that the mutation was segregating with the phenotype, and previously described functional studies reported the variant to be pathogenic through impaired binding of a transcription factor (Li *et al* 2016). Predictive testing for this variant is now being established in the clinical diagnostic setting for extended family members.

I sought to determine the cause of the reduced gene transcription in the remaining 3 samples. However, in addition to apparently normal UDS of the entire genomic *APC* locus, including the promoter regions and UTRs, karyotype analysis and promoter methylation studies also gave normal results. We are planning that these patients and appropriate members of their families undergo whole genome sequencing with the aim of searching for variants in the regions of the genome which have not yet been examined.

UDS only identified one pathogenic mutation - the *APC* 5'UTR variant in Halo46 described above. There was no evidence for any likely pathogenic variants outside of the ORF or for *APC* mosaicism in the other patients. With regards to *APC* mosaicism, the average depth of coverage of *APC* using the HiSeq 2500 was 1665.5x, so if there were any cases with significant levels of mutant alleles in blood DNA, they were unlikely to have been missed. However, others have published evidence indicating that somatic mosaicism restricted to the gut or a section of the bowel may be relatively common (Jansen *et al* 2017; Spier *et al* 2016), limiting the usefulness of studying DNA derived from blood.

6.1.1 Strengths and Weaknesses of the Study of *APC* and *MUTYH*

There were several strengths to this study. DNA was obtained for all study participants, and RNA from the majority of individuals. FFPE tissue was available for most cases when it was required for IHC analysis or for somatic DNA analysis. Furthermore, it was possible to access patients and their medical notes when they were registered with the AWMGS. We worked closely with the WGP, which allowed access to NGS technologies, making it possible to examine whole genomic loci and to perform exome sequencing.

One of the key problems encountered throughout this project was the lack of adequate clinical information about the patients. Often the phenotype, histopathological assessment of lesions and family history were not complete, so it was difficult to determine whether cases were familial or sporadic, and what the true polyp counts/ morphology were. We found that despite acquiring written patient consent to access medical records under an ethically approved study, it was often difficult to access clinical details for patients who were not under follow up with the regional genetics service in Wales. Busy hospital services have priorities other than supporting research studies and the dispersion of clinical data in the NHS makes its retrieval time consuming. The gradual move to digitalised records will likely benefit clinical research in the longer term. The optimal information would have included a complete family history, the age of onset of polyposis, the cumulative numbers of polyps, the size and location of polyps and their microscopic morphology.
It also often proved difficult to recruit family members of the proband. This made analysis difficult as it was often not possible to ascertain whether identified variants were occurring *de novo* or were segregating with disease if multiple family members were affected.

Part of the study involved using NGS technologies to carry out UDS. NGS generates large amounts of data, and a challenge is to identify what is clinically significant. In this thesis, CADD scoring was used as a tool to indicate the likelihood of pathogenicity and to therefore prioritise variants for further study. Although the CADD score gives a measure of predicted deleteriousness, this is not the same as its likelihood of causing a clinical phenotype. Furthermore, CADD scoring was used to appraise all variants. Recent evidence suggests that ‘CADD scores may be useful for identification of pathogenic intronic or nonsynonymous variants in

targeted testing situations when used in combination with other data. However our data suggest that CADD scores are unlikely to be useful for identifying disease-causing mutations in other noncoding regions in cancer-risk genes' (Mather C. *et al* 2016), so the utility of this method in scoring variants in regions such as promoters/ UTRs is questionable. The authors acknowledge that there is a distinction between variants that are functionally deleterious and clinically pathogenic (Mather C. *et al* 2016). They note that there are many situations in which a deleterious variant does not cause clinical phenotype (Mather C. *et al* 2016).

6.1.2 Clinically Translatable Outcomes

The key clinically translatable outcomes from this stage of the project are detailed below:

1. ***APC* promoter sequencing and sequencing of Exon 1/ 5'UTR.** In the clinical diagnostic setting, *APC* and *MUTYH* undergo sequence and dosage analysis through sequencing and MLPA respectively. Standard protocols involve sequencing of the coding exons (*APC* and *MUTYH*) and dosage analysis of coding exons (*APC* and *MUTYH*) and promoter (*APC* only) (2.2.3.1). It is recommended that diagnostic laboratories consider including the *APC* promoters 1A and 1B and Exon 1 in protocols for patients who remain NMI after sequencing of coding exons. This would allow the identification of variants which could result in abnormalities of gene expression.
2. **RNA studies.** For NMI patients, diagnostic laboratories could consider RNA studies as part of their investigative tool kit. cDNA sequencing of *APC* is a relatively simple procedure to undertake. If abnormalities were identified, sequencing of the entire genomic locus of *APC* could be undertaken to determine the underlying cause. Quantitative studies of *APC* transcription could also be considered. However, this would first necessitate a large-scale study to determine the normal range of *APC* transcription in a healthy cohort and to compare this to expression levels in known *APC* mutation carriers. qPCR was used in this study to identify reduced *APC* transcription in 4 of 45 NMI patients when compared to a healthy control cohort of 40 individuals, but for clinical applications the control cohort would need to be

much larger, and multiple Taqman probes would need to be used. Again, if an individual was found to have reduced *APC* transcription, sequencing of the entire genomic locus could be performed to try to determine its cause. However, as was seen in this work, the cause isn't always apparent. Through research such as the 100 000 Genomes Project, which includes a cohort of patients with multiple bowel polyps who were NMI following testing in the NHS, further clinically significant variants in gene enhancers/ upstream regions may be characterised.

6.2 Targeted and Whole Exome Sequencing

Targeted exome sequencing was initially carried out to interrogate 15 genes known to be, or possibly, involved in the pathogenesis of colorectal neoplasia. A total of 33 variants across 32 patients were selected for further investigation: 22 of these were validated. The variants which were not validated were largely those occurring at a low frequency or in regions with low NGS coverage. Whole exome sequencing (WES) was undertaken on a subset of 24 patients, based on their phenotype, the availability of relatives and those patients in whom a novel and possibly causative mutation had already been identified, to confirm that missed mutations in the known genes were not involved. 106 variants were selected for validation, and 100 of these were validated. Again, the variants which were not validated were largely present in regions of low NGS coverage.

The genes/ variants which were selected in the short term for functional characterisation were those identified as part of the targeted exome testing: variants in *AXIN2, POLE*, *POLD1* and the MMR genes. Although many other validated variants were potentially interesting, they were not taken forward for further studies. The reasons for not selecting other variants included 1) lack of availability of relatives, 2) the variant was in a gene already associated with a disease phenotype other than colorectal neoplasia, 3) it was too frequent to be a reasonable candidate for the relatively uncommon phenotype of polyposis or 4) there was not sufficient evidence of a likely role in colorectal neoplasia to prioritise the gene concerned. As mentioned previously, some variants were identified which were known to be clinically significant in contexts other than colorectal neoplasia. Where these were clinically actionable, they were fed back to the patient's clinician, according to the study protocol.

6.2.1 Strengths and Weaknesses of Targeted and Whole Exome Sequencing in NMI Polyposis Patients

As described above (6.1.1) the strengths to this study included the availability of DNA and RNA and the access to NGS technologies. A further strength is that we were fortunate to have had input from international experts, Professor Seruca, and industry, AmbryGen, to assist in variant appraisal.

The most important factors limiting progress in the search for novel variants associated with polyposis were the lack of additional family members and the size of the study cohort. Other studies have been more successful because they were larger or prioritised the study of familial cases (Adam *et al* 2016; Weren *et al* 2015) with availability of samples from multiple affected relatives.

Another issue faced was variable coverage with WES. The mean depth of coverage across the 29 patients/ relatives was 61.99 reads, but this varied from 14.02 reads (Halo66) to 139.76 reads (Halo61). Furthermore, the percentage of the target region covered at 50x depth was only 43.9%. When variants were not validated following Sanger sequencing, these tended to have occurred in regions of low coverage.

Exome sequencing generates large amounts of data, including many novel variants, leading to a challenge in the selection of variants for further study. CADD scores were one tool used for variant prioritisation in the current work, and their limitations have already been discussed (6.1.1). Ultimately choices have to be made based upon a variant's predicted impact on gene function, previous literature on the functions of the gene involved and genetic data, if available, on *de novo* occurrence or segregation with the disease phenotype.

The work in this part of the project required database interrogation to aid in prioritising identified variants. As described in Chapter 4 (4.4.1) it is apparent that genetic variants identified through clinical research projects are often not recorded in publicly accessible databases. It would be excellent if *all* variants were recorded in a centralised repository, such as LOVD, as this would allow improved understanding of disease pathogenesis and therefore improved patient care. With the advent of the 100 000 Genomes Project it is hoped that there will be improved understanding of the role of the genome in health and disease. The project seeks

to enhance the clinical interpretation of the data and derive new findings from the data (https://www.genomicsengland.co.uk/faqs-about-gecip/ accessed 10/5/2018).

6.2.2 Recommendations Resulting from Exome Sequencing

1. **Ensure attempts are made to recruit family members at the onset of the study.** Affected and/ or unaffected family members are required in order to meaningfully evaluate variants identified in probands. I would recommend that for future genetic research similar to this, family members should be recruited to studies at their onset, if at all possible.
2. **Record all variants in a pubic database.** The aim of all clinical research should be to improve patient care. We were very fortunate throughout this study to have had input from world experts and from Industry, for example AmbryGen. The optimal situation would be to record all variants identified through genetic research in a publicly-available repository, as described in sections 4.4.1 and 6.2.1.

6.3 Functional Characterisation of Variants Identified in the Mismatch Repair Genes, *POLE*, *POLD1* and *AXIN2*

The final phase of this project sought to further characterise variants in the MMR genes, *POLE* and *POLD1* and *AXIN2* to gain evidence either for or against their clinical significance.

6.3.1 The MMR Genes

Three patients were found to carry germline variants in one of the MMR genes (Halo26, Halo45 and Halo70). Tumour samples were available for all three patients. The pathogenic effects of the mutations were examined through IHC analysis of the MMR proteins and MSI testing of tumour DNA. In all cases, MMR IHC was normal, and all lesions tested were microsatellite stable. This lead to the conclusion that the variants are unlikely to be pathogenic. This in keeping with their prior probability of pathogenicity scores (5.2.6.1) which were all <0.95.

6.3.2 *POLE* and *POLD1*

Four patients were found to carry germline mutations in *POLE* or *POLD1* (Halo15, Halo18, Halo27 and Halo61). Basic *in silico* and functional studies were initially carried out. These included structural modelling to try and predict whether the mutations could theoretically impair the functioning of the proteins and studies on tumour samples to assess their MSI status and to test for the somatic 'hot spot' mutations which have been reported in tumours from carriers of pathogenic *POLE* mutations (Palles *et al* 2013). Based on structural modelling, the p.Glu2140Lys mutation in Halo18 was deemed unlikely to have functional consequences so was not taken further forward.

The 'gold standard' of variant assessment was to determine the 'mutation signature' of the tumours arising in carriers of mutations in the *pol* genes. Known pathogenic mutations impair the proofreading activity of the polymerases, and they are associated with a hypermutated genetic signature in tumours. Somatic genetic analysis was performed in collaboration with Professor Ian Tomlinson and Dr. Claire Palles (Oxford University/ Birmingham University). SmMIPs were used to capture 30 genes. Unfortunately, sequencing was not successful for Halo61, so p.Tyr623Cys remains a VUS. There was a seven-month delay between preparing the samples for sequencing and obtaining the initial results from sequencing. This was due to factors beyond our control (the relocation of the Tomlinson group from Oxford to Birmingham). However, it has meant that we currently only have the raw sequencing results for Halo15, for which bioinformatic analysis is awaited. The tumour sequencing results for Halo27 are pending. Therefore, at the present time, the variants identified in Halo15 and Halo27, c.1411A>G:p.Met471Val and c.2861C>T:p.Thr954Met still remain as VUS.

6.3.3 *AXIN2* c.1642G>T:p.Glu548*

Two sisters participating in this study were found to carry a germline mutation in *AXIN2,* c.1642G>T:p.Glu548*. When further family members were recruited, this was confirmed to segregate with occurrence of colorectal neoplasia in an AD manner. All individuals affected by colorectal neoplasia also had a history of oligodontia, and there were variable signs of ectodermal dysplasia. These findings are in keeping with previous reports in the literature describing *AXIN2* mutations in the context of colorectal polyposis (Marvin *et al* 2011; Lammi *et al* 2004). Further

work investigating this mutation confirmed that it was present in patient-derived RNA samples and that it was capable of producing a protein approximately 30k DA smaller than the wild-type protein. The protein appeared to be more stable then wild-type protein, again in keeping with the previous literature (Mazzoni *et al* 2015). Assessment of the mutation's impact on Wnt-pathway activation was investigated using the luciferase assay in HEK293 TCF-Luc cell lines. Results showed that Short-*AXIN2,* WT/ Short-*AXIN2* and Fearon-*AXIN2* all had impaired ability to inhibit Wnt-pathway activation, although this was only statistically significant for the WT/ Short-*AXIN2* mix.

To further investigate the mutation, β-catenin IHC was performed on tumours arising in carriers of germline *AXIN2* mutations. There was evidence of nuclear β-catenin, but only in a very small proportion of tumours, and it was more overt in higher-grade lesions. This suggested that there is Wnt-pathway activation in colorectal tumours from the *AXIN2* mutation carriers investigated, but that it may be occurring at a later stage of tumourigenesis compared to sporadic or FAP-associated lesions and/ or it may be a subtler effect.

I attempted to screen the *APC* MCR to determine whether somatic mutations in *APC* were present, as this is a common finding in FAP-associated and sporadic colorectal neoplasia. Unfortunately, the quality and quantity of DNA extracted from tumours was poor, and MCR sequencing was incomplete. Nonetheless, there appeared to be fewer somatic *APC* mutations than would have been expected in the adenomas, although this was not confirmed statistically due to the lack of a robust control set. This raises the possibility that there may be a novel genetic pathway of tumourigenesis in carriers of germline *AXIN2*-mutations, such that *APC* mutations are not required for the initial stages of tumourigenesis. This will be further investigated through WES of tumour DNA.

To summarise, we have identified a family in which a truncating *AXIN2* mutation is segregating with colorectal neoplasia and oligodontia in an AD manner. It is not clear whether the additional signs of ectodermal dysplasia, thin hair and nail abnormalities, are part of the phenotype or are incidental findings. The results obtained from *in vitro* studies are in keeping with the effects of the mutation at least being partially mediated through activation of the canonical Wnt-signalling pathway. Interestingly, a key extra-intestinal finding appears to be oligodontia. The role of the Wnt-pathway in tooth formation has been discussed in Chapter 5 (5.4.1).

The results of our study are consistent with previous reports in the literature – notably that truncating *AXIN2* mutations have been found to segregate with colorectal neoplasia and oligodontia in an AD manner (Marvin *et al* 2011; Lammi *et al* 2004). The colorectal phenotype was variable in our family, like that reported in other families. The polyp count is variable, the presence or absence of CRC is variable, and the morphology of the lesions present is variable: both adenomas and HPPs/ serrated lesions have been described. In this study, colorectal tumours arising in the context of a germline *AXIN2* mutation were characterised by IHC and molecular genetic analysis. This pointed to potential differences from sporadic or FAP-related tumours. However, further characterisation is required to confirm or refute these differences. It is possible that adenomas may result from activation of the Wnt-signalling pathway, whereas HPPs/ serrated lesions may be the result of effects on other signalling pathways impacted by *AXIN2,* for example through Ras/ ERK signalling pathways or SAPK/ JNK pathways (Mazzoni and Fearon 2014).

There is mounting evidence for the existence of an *AXIN2*-Associated Polyposis Syndrome. Although it is likely to be rare, its true prevalence may be underestimated due to the variable GI phenotype. It is possible that *AXIN2* mutations are highly penetrant with regards to the dental phenotype but are of variable penetrance in terms of the colorectal phenotype. To further characterise the syndrome, it might be sensible to aim to identify patients through dental clinics rather than through gastroenterology/ colorectal surgery.

6.4 Thesis Conclusions

The aim of this study was to identify novel genes/ mechanisms of disease in a cohort of patients with colorectal polyposis who had not had a mutation identified in a clinical genetics diagnostic setting.

A total of 60 NMI polyposis patients were included in this project. Putative pathogenic mutations/ disease mechanisms have been identified in 8 of these patients: 4 with evidence of reduced *APC* expression, in one of whom the underlying mutation has been confirmed (Halo46); 2 with variants in *POLE/ POLD1* in whom further analysis of adenomas is in progress, and 2 with mutations in *AXIN2.*

The phenotype of the patients recruited to this study was highly variable, in terms of the age of patients, the family histories, the polyp counts and morphologies. There are several possible explanations as to why we did not identify mutations in more of the patients: variants may be in genomic regions not yet interrogated; they may be restricted to a tissue not examined, for example the gut epithelium; epigenetic factors may have a role; some patients may have their phenotype as a result of oligogenic inheritance and some patients may have numerous sporadic lesions and represent one end of a normal distribution of sporadic lesions in the non-predisposed population.

There are clinical recommendations that result from this project, namely that *APC* promoter/ Exon 1 screening is carried out for patients who remain NMI after standard genetic testing protocols and that RNA studies are considered for NMI patients. With further characterisation it is possible that a case for *AXIN2* to be included on polyposis/ CRC gene testing panels will be made.

Further work which will be done following on from this project will be to carry out whole genome sequencing for those patients with evidence of reduced *APC* transcription and to examine the genetic signature of tumours developing in carriers of *AXIN2* mutations.

The ultimate aim of this work is to optimise patient care and to prevent the development of CRC. Once causative mutations are identified, family members can be screened, and appropriate clinical management instituted. The work is already having an impact on the family of Halo46 and will hopefully have a positive impact on other families and future generations. A relative of Halo46 has written:

'I would also like to thank you for the information you passed to … This means my son will be able to be checked for FAP without the initial endoscopy. Just this small advance has lifted a world of anxiety from my shoulders, thank you so very much indeed.'

References

Adam, R. *et al* (2016). Exome Sequencing Identifies Biallelic MSH3 Germline Mutations as a Recessive Subtype of Colorectal Adenomatous Polyposis. *Am J Hum Genet* 99(2), pp. 337-351.

Al-Tassan, N. *et al* (2002). Inherited variants of MYH associated with somatic G:C-->T:A mutations in colorectal tumors. *Nat Genet* 30(2), pp. 227-232.

Albuquerque, C. *et al* (2002). The 'just-right' signalling model: APC somatic mutations are selected based on a specific level of activation of the beta-catenin signalling cascade. *Hum Mol Genet* 11(13), pp. 1549-1560.

Aretz S. *et al* (2014). MUTYH-associated polyposis (MAP): evidence for the origin of the common European mutations p.Tyr179Cys and p.Gly396Asp by founder events. *Eur J Hum Genet,* 22(7), pp. 923-9.

Aretz, S. *et al* (2007). Somatic APC mosaicism: a frequent cause of familial adenomatous polyposis (FAP). *Hum Mutat* 28(10), pp. 985-992.

Arminski, TC. and Mclean, DW. (1964). incidence and distribution of adenomatous polyps of the colon and rectum based on 1,000 autopsy examinations. *Dis Colon Rectum* 7, pp. 249-261.

Arnold, CN. *et al* (2004). APC promoter hypermethylation contributes to the loss of APC expression in colorectal cancers with allelic loss on 5q. *Cancer Biol Ther* 3(10), pp. 960-964.

Baker, SJ. *et al* (1989). Chromosome 17 deletions and p53 gene mutations in colorectal carcinomas. *Science* 244(4901), pp. 217-221.

Bellido, F. *et al* (2016). POLE and POLD1 mutations in 529 kindred with familial colorectal cancer and/ or polyposis: review of reported cases and recommendations for genetic testing and surveillance. *Genet Med* 18(4), pp. 325-332.

Berdasco, M. and Esteller, M. (2010). Aberrant epigenetic landscape in cancer: how cellular identity goes awry. *Dev Cell* 19(5), pp. 698-711.

Bodmer, W. (1999). Familial adenomatous polyposis (FAP) and its gene, APC. *Cytogenet Cell Genet* 86, pp. 99–104.

Boland, CR. and Goel, A. (2010) Microsatellite instability in colorectal cancer. *Gastroenterology* 138(6), pp. 2073-2087.e3.

Bollag, G. *et al* (2012). Vemurafenib: the first drug approved for BRAF-mutant cancer. *Nat Rev Drug Discov* 11(11), pp. 873-886.

Boparai, KS. *et al* (2008). Hyperplastic polyps and sessile serrated adenomas as a phenotypic expression of MYH-associated polyposis. *Gastroenterology* 135(6), pp. 2014-2018.

Bos, JL. *et al* (1987). Prevalence of ras gene mutations in human colorectal cancers. *Nature* 327(6120), pp. 293-297.

Bolton, A. (2016). Genetic mechanisms in colorectal polyposis. *Intercalated year thesis.*

Bradford, B. (2017). Use of Whole Exome Sequencing for the Identification of Novel Genetic Mechanisms in Colorectal Polyposis. *Intercalated year thesis.*

Brink, M. *et al* (2003). K-ras oncogene mutations in sporadic colorectal cancer in The Netherlands Cohort Study. *Carcinogenesis* 24(4), pp. 703-710.

Buchanan, DD. *et al* (2017). Lack of evidence for germline RNF43 mutations in patients with serrated polyposis syndrome from a large multinational study. *Gut* 66(6), pp. 1170-1172.

Burt, R. (2007). Inheritance of Colorectal Cancer. *Drug Discov Today Dis Mech* 4(4), pp. 293-300.

Bülow, S. *et al* (2006). The history of familial adenomatous polyposis. *Fam Cancer* 5(3), pp. 213-220.

Calva D. and Howe J (2009) Juvenile Polyposis. In: *Cancer Syndromes*. Ed Riegert-Johnson, DL *et al* Pub Bethesda (MD): National Center for Biotechnology Information (US).

Cancer Research UK (2015). Bowel cancer statistics. Available at http://www.cancerresearchuk.org/health-professional/cancer-statistics/statistics-by-cancer-type/bowel-cancer#heading-Zero (accessed 04/12/2017).

Castellsagué E. *et al* (2010). Allele-specific expression of APC in adenomatous polyposis families. *Gastroenterology* 139(2), pp. 439-447.

Chapman, I. (1963). Adenomatous polypi of large intestine: incidence and distribution. *Ann Surg* 157, pp. 223-226.

Charames, GS. *et al* (2008). A large novel deletion in the APC promoter region causes gene silencing and leads to classical familial adenomatous polyposis in a Manitoba Mennonite kindred. *Hum Genet* 124(5), pp. 535-541.

Cheadle, JP. and Sampson, JR. (2007). MUTYH-associated polyposis--from defect in base excision repair to clinical genetic testing. *DNA Repair (Amst)* 6(3), pp. 274-279.

Chen, X. *et al* (2008). Allelic imbalance in BRCA1 and BRCA2 gene expression is associated with an increased breast cancer risk. *Hum Mol Genet.* 17(9), pp. 1336-1348.

Chung, SJ. *et al* (2010). Prevalence and risk of colorectal adenoma in asymptomatic Koreans aged 40-49 years undergoing screening colonoscopy. *J Gastroenterol Hepatol* 25(3), pp. 519-525.

Cleary, SP. *et al* (2008). Missense polypmophisms in the *Adenomatous Polyposis Coli* gene and colorectal cancer risk. *Dis Colon Rectum* 51(10), pp. 1467-1474.

Conne, B. *et al.* (2000). The 3' untranslated region of messenger RNA: A molecular 'hotspot' for pathology? *Nat Med* 6(6), pp. 637-641.

Correa, P. *et al* (1977). The epidemiology of colorectal polyps: prevalence in New Orleans and international comparisons. *Cancer* 39(5), pp. 2258-2264.

Cottrell, S. *et al* (1992). Molecular analysis of APC mutations in familial adenomatous polyposis and sporadic colon carcinomas. *Lancet* 340(8820), pp. 626-630.

Croitoru, ME. *et al* (2007). Germline MYH mutations in a clinic-based series of Canadian multiple colorectal adenoma patients. *J Surg Oncol* 95(6), pp. 499-506.

Cybulski, C. *et al* (2004). CHEK2 is a multiorgan cancer susceptibility gene. *Am J Hum Genet* 75(6), pp. 1131-1135.

Cybulski, C. *et al* (2007). Germline CHEK2 mutations and colorectal cancer risk: different effects of a missense and truncating mutations? *Eur J Hum Genet* 15(2), pp. 237-241.

Das, S. *et al.* (2015). Human Epidermal Growth Factor Receptor 2 (HER2) Impedes MLK3 Kinase Activity to Support Breast Cancer Cell Survival. *The Journal of Biological Chemistry* 290, pp. 21705–21712.

David, SS. *et al* (2007). Base-excision repair of oxidative DNA damage. *Nature* 447(7147), pp. 941-950.

Davies, A. (2015). Genetic mechanisms in colorectal polyposis. *Intercalated year thesis.*

De Benedetti L. *et al* (1994). Association of APC gene mutations and histological characteristics of colorectal adenomas *Cancer Research* 54, pp. 3553-3556.

Delano, WL. (2002). The PyMOL molecular graphics system. Delano Scientific LLC, Palo Alto *http://www.pymol.org.*

DiSario, JA. *et al* (1991). Prevalence and malignant potential of colorectal polyps in asymptomatic, average-risk men. *Am J Gastroenterol* 86(8), pp. 941-945.

Dolwani, S. *et al* (2007). Analysis of inherited MYH/ (MutYH) mutations in British Asian patients with colorectal cancer. *Gut* 56(4), p. 593.

Emsley, P. and Cowtan, K. (2004). Coot: model-building tools for molecular graphics. *Acta Crystallogr D Biol Crystallogr* 60(Pt 12 Pt 1), pp. 2126-2132.

Enholm, S. *et al* (2003). Proportion and phenotype of MYH-associated colorectal neoplasia in a population-based series of Finnish colorectal cancer patients. *Am J Pathol* 163(3), pp. 827-832.

Esteller, M. *et al.* (2000). Analysis of adenomatous polyposis coli promoter hypermethylation in human cancer. *Cancer Res* 60(16), pp. 4366-4371.

Farrington, SM. *et al* (2005). Germline susceptibility to colorectal cancer due to base-excision repair gene defects. *Am J Hum Genet* 77(1), pp. 112-119.

Fearnhead, NS. *et al* (2001). The ABC of APC. *Hum Mol Genet* 10(7), pp. 721-733.

Fearnhead, NS. *et al* (2002). Genetics of colorectal cancer: hereditary aspects and overview of colorectal tumorigenesis. *Br Med Bull* 64, pp. 27-43.

Fearon, ER. and Vogelstein, B. (1990). A genetic model for colorectal tumorigenesis. *Cell* 61(5), pp. 759-767.

Fleming, NI. *et al* (2013). SMAD2, SMAD3 and SMAD4 mutations in colorectal cancer. *Cancer Res* 73(2), pp. 725-735.

Fodde, R. (2002). The APC gene in colorectal cancer. *Eur J Cancer* 38(7), pp. 867-871.

Forrester, K. *et al* (1987). Detection of high incidence of K-ras oncogenes during human colon tumorigenesis. *Nature* 327(6120), pp. 298-303.

Forsberg, AM. *et al* (2012). Prevalence of colonic neoplasia and advanced lesions in the normal population: a prospective population-based colonoscopy study. *Scand J Gastroenterol* 47(2), pp. 184-190.

Gala, MK. *et al* (2014). Germline mutations in oncogene-induced senescence pathways are associated with multiple sessile serrated adenomas. *Gastroenterology* 146(2), pp. 520-529.

Giacosa, A. *et al* (2004). Epidemiology of colorectal polyps. *Tech Coloproctol* 8 Suppl 2, pp. s243-247.

Giardiello, FM. *et al* (2014). Guidelines on genetic evaluation and management of Lynch syndrome: a consensus statement by the US Multi-Society Task Force on colorectal cancer. *Am J Gastroenterol* 109(8), pp. 1159-1179.

Giardiello, FM. *et al* (1991). Colorectal neoplasia in juvenile polyposis or juvenile polyps. *Arch Dis Child* 66(8), pp. 971-975.

Gismondi, V. *et al.* (2004). Prevalence of the P.TYR165CYS, P.GLY382ASP and 1395delGGA germline mutations of the MYH gene in Italian patients with adenomatous polyposis coli and colorectal adenomas. *Int J Cancer* 109(5), pp. 680-684.

Gonzalez, KD. *et al* (2009). Beyond Li Fraumeni Syndrome: clinical characteristics of families with p53 germline mutations. *J Clin Oncol* 27(8), pp. 1250-1256.

Grady, WM. and Carethers, JM. (2008). Genomic and epigenetic instability in colorectal cancer pathogenesis. *Gastroenterology* 135(4), pp. 1079-1099.

Groden, J. *et al* (1993). Mutational analysis of patients with adenomatous polyposis: identical inactivating mutations in unrelated individuals. *Am J Hum Genet* 52(2), pp. 263-272.

Groden, J. *et al*. (1991). Identification and characterization of the familial adenomatous polyposis coli gene. *Cell* 66(3), pp. 589-600.

Groen, EJ. *et al* (2008). Extra-intestinal manifestations of Familial Adenomatous Polyposis. *Annals of Surg Onc* 15(9), pp. 2439-2450.

Guarinos, C. *et al* (2012). Serrated polyposis syndrome: molecular, pathological and clinical aspects. *World J Gastroenterol* 18(20), pp. 2452-2461.

Guinney, J. *et al* (2015). The consensus molecular subtypes of colorectal cancer. *Nat Med* 21(11), pp. 1350-1356.

Hahn, SA. *et al* (1996). DPC4, a candidate tumor suppressor gene at human chromosome 18q21.1. *Science* 271(5247), pp. 350-353.

Hanahan, D. and Weinberg, RA. (2000). The hallmarks of cancer. *Cell* 100(1), pp. 57-70.

Hanahan, D. and Weinberg, RA. (2011). Hallmarks of cancer: the next generation. *Cell* 144(5), pp. 646-74.

Hanssen, AM. and Fryns, JP. (1995). Cowden syndrome. *J Med Genet* 32(2), pp. 117-119.

Hao, X. *et al* (1997). Reciprocity between membranous and nuclear expression of beta-catenin in colorectal tumours. *Virchows Arch* 431(3), pp. 167-172.

Hardwick, JC. *et al* (2008). Bone morphogenetic protein signalling in colorectal cancer. *Nat Rev Cancer* 8(10), pp. 806-812.

Harewood, L. and Fraser, P. (2014). The impact of chromosomal rearrangements on regulation of gene expression. *Hum Mol Genet* 23(R1), pp. R76-82.

Harrison-Cripps, W. (1882). Two cases of disseminated polypus of the rectum. *Transactions of the Pathological Society of London* 33, pp. 165-168.

Hertz, AF. (1914). Four Cases of Rectal Polypus occurring in One Family. *Proc R Soc Med* 7(Surg Sect), pp. 255-256.

Hes, FJ. *et al* (2008). Somatic APC mosaicism: an underestimated cause of polyposis coli. *Gut* 57(1), pp. 71-76.

Hizawa, K. *et al* (1993). Cancer in Peutz-Jeghers syndrome. *Cancer* 72(9), pp. 2777-2781.

Hogg, M. *et al* (2014). Structural basis for processive DNA synthesis by yeast DNA polymerase ε. *Nat Struct Mol Biol* 21(1), pp. 49-55.

Howe, JR. *et al* (2007). *ENG* mutations in *MADH4/BMPR1A* mutation negative patients with juvenile polyposis. *Clinical Genetics,* 71(1), pp.91-92.

Hughes, LA. *et al* (2012). The CpG island methylator phenotype in colorectal cancer: progress and problems. *Biochim Biophys Acta* 1825(1), pp. 77-85.

Hyman, NH. *et al* (2004). Hyperplastic polyposis and the risk of colorectal cancer. *Dis Colon Rectum* 47(12), pp. 2101-2104.

Iacopetta, B. (2003). TP53 mutation in colorectal cancer. *Hum Mutat* 21(3), pp. 271-276.

Iino, H. *et al* (2000). DNA microsatellite instability and mismatch repair protein loss in adenomas presenting in hereditary non-polyposis colorectal cancer. *Gut* 47(1), pp. 37-42.

Isidro, G. *et al* (2004). Germline MUTYH (MYH) mutations in Portuguese individuals with multiple colorectal adenomas. *Hum Mutat* 24(4), pp. 353-354.

Iwamoto, M. *et al* (2000). Expression of beta-catenin and full-length APC protein in normal and neoplastic colonic tissues. *Carcinogenesis* 21(11), pp. 1935-1940.

Jaeger, E. *et al*. (2012). Hereditary mixed polyposis syndrome is caused by a 40-kb upstream duplication that leads to increased and ectopic expression of the BMP antagonist GREM1. *Nat Genet* 44(6), pp. 699-703.

Jansen, AM. *et al* (2017). Distinct patterns of somatic mosaicism in the APC gene in neoplasms from patients with unexplained adenomatous polyposis. *Gastroenterology* 152(3), pp. 546-549.e543.

Järvinen, E. *et al* (2018). Mesenchymal Wnt/ β-catenin signalling limits tooth number. *Development* 145, pp. 1-12.

Jass, JR. and Burt, R. (2000). Hyperplastic Polyposis. *World Health Organisation Classification of Tumours, Pathology and Genetics of Tumours of the Digestive System.* Ed. Hamilton, SR. and Aaltonen, LA. Pub. IARC Press, Lyon.

Jegers, H. *et al* (1949). Generalized intestinal polyposis and melanin spots of the oral mucosa, lips and digits; a syndrome of diagnostic significance. *N Engl J Med* 241(26), pp. 1031-1036.

Jenne, DE. *et al* (1998). Peutz-Jeghers syndrome is caused by mutations in a novel serine threonine kinase. *Nat Genet* 18(1), pp. 38-43.

Jones, PA. and Baylin, SB. (2007). The epigenomics of cancer. *Cell* 128(4), pp. 683-692.

Jones, S. *et al* (2002). Biallelic germline mutations in MYH predispose to multiple colorectal adenoma and somatic G:C-->T:A mutations. *Hum Mol Genet* 11(23), pp. 2961-2967.

Joslyn, G. *et al* (1991). Identification of deletion mutations and three new genes at the familial polyposis locus. *Cell* 66(3), pp. 601-613.

Järvinen, H. and Franssila, KO. (1984). Familial juvenile polyposis coli; increased risk of colorectal cancer. *Gut* 25(7), pp. 792-800.

Kadiyska, T. *et al* (2013). APC promoter 1B deletion in familial polyposis-implications for mutation-negative families. *Clin Genet* 85(5), pp. 452-457.

Kairupan, CF. *et al* (2005). Mutation analysis of the MYH gene in an Australian series of colorectal polyposis patients with or without germline APC mutations. *Int J Cancer* 116(1), pp. 73-77.

Kalady, MF. *et al* (2011). Defining phenotypes and cancer risk in hyperplastic polyposis syndrome. *Dis Colon Rectum* 54(2), pp. 164-170.

Kanter-Smoler, G. *et al* (2006). Novel findings in Swedish patients with MYH-associated polyposis: mutation detection and clinical characterization. *Clin Gastroenterol Hepatol* 4(4), pp. 499-506.

Kaschula, RO. (1971). Mixed juvenile, adenomatous and intermediate polyposis coli: report of a case. *Dis Colon Rectum* 14(5), pp. 368-374.

Keller, G. *et al* (2004). Germline mutations of the E-cadherin(CDH1) and TP53 genes, rather than of RUNX3 and HPP1, contribute to genetic predisposition in German gastric cancer patients. *J Med Genet.* 41(6), pp. e89.

Kilpivaara, O. *et al* (2006). CHEK2 I157T associates with familial and sporadic colorectal cancer. *J Med Genet* 43(7), pp. e34.

Kim, HC. *et al* (2000). The E-cadherin gene (CDH1) variants T340A and L599V in gastric and colorectal cancer patients in Korea. *Gut* 47(2), pp. 262-267.

Kinzler, KW. *et al*. (1991). Identification of FAP locus genes from chromosome 5q21. *Science* 253(5020), pp. 661-665.

Knudson, AG. (1971). Mutation and cancer: statistical study of retinoblastoma. *Proc Natl Acad Sci U S A* 68(4), pp. 820-823.

Kobayashi, M. *et al* (2000). Nuclear translocation of beta-catenin in colorectal cancer. *Br J Cancer* 82(10), pp. 1689-1693.

Küry, S. *et al* (2007). The thorough screening of the MUTYH gene in a large French cohort of sporadic colorectal cancers. *Genet Test* 11(4), pp. 373-379.

Lage, P. *et al* (2004). Management of Portuguese patients with hyperplastic polyposis and screening of at-risk first-degree relatives: a contribution for future guidelines based on a clinical study. *Am J Gastroenterol* 99(9), pp. 1779-1784.

Laken, SJ. *et al* (1999). Analysis of masked mutations in familial adenomatous polyposis. *Proc Natl Acad USA* 96(5), pp. 2322-2326.

Lam-Himlin, D. *et al* (2014). Gastric polyps and polyposis syndromes. *Diag Histopath* 20(1), pp. 1-11.

Lammi, L. *et al* (2004). Mutations in AXIN2 cause familial tooth agenesis and predispose to colorectal cancer. *Am J Hum Genet* 74(5), pp. 1043-1050.

Leggett, B. and Whitehall, V. (2010). Role of the serrated pathway in colorectal cancer pathogenesis. *Gastroenterology* 138(6), pp. 2088-2100.

Lejeune, S. *et al* (2006). Low frequency of AXIN2 mutations and high frequency of MUTYH mutations in patients with multiple polyposis. *Hum Mutat* 27(10), p. 1064.

Li, J. *et al* (2016). Point mutations in exon 1B of APC reveal gastric adenocarcinoma and proximal polyposis of the stomach as familial adenomatous polyposis variant. *Am J Hum Gen* 98(5), pp. 830-842.

Lin, OS. *et al* (2006). Screening colonoscopy in very elderly patients: prevalence of neoplasia and estimated impact on life expectancy. *JAMA* 295(20), pp. 2357-2365.

Lipton, L. *et al.* 2003a. Carcinogenesis in MYH-associated polyposis follows a distinct genetic pathway. *Cancer Res* 63(22), pp. 7595-7599.

Lipton, L. *et al* (2003b). Germline mutations in the TGF-beta and Wnt signalling pathways are a rare cause of the "multiple" adenoma phenotype. *J Med Genet* 40(4), p. e35.

Lloyd, KM. and Dennis, M. (1963). Cowden's disease. A possible new symptom complex with multiple system involvement. *Ann Intern Med* 58, pp. 136-142.

Lockhart-Mummery P. (1925). Cancer and Heredity. *Lancet* i, pp.4270429.

Logan, CY. and Nusse, R. (2004). The Wnt signalling pathway in development and disease. *Annu Rev Cell Dev Biol* 20, pp. 781-810.

Lubbe, SJ. *et al* (2009). Clinical implications of the colorectal cancer risk associated with MUTYH mutation. *J Clin Oncol* 27(24), pp. 3975-3980.

Lynch, HT. *et al* (2015). Milestones of Lynch syndrome: 1895-2015. *Nat Rev Cancer* 15(3), pp. 181-194.

Marvin, ML. *et al* (2011). AXIN2-associated autosomal dominant ectodermal dysplasia and neoplastic syndrome. *Am J Med Genet A* 155A(4), pp. 898-902.

Mather, C. *et al* (2016). CADD score has limited clinical validity for the identification of pathogenic variants in noncoding regions in a hereditary cancer panel. *Genetics in Medicine* (18), pp. 1269-1275.

Mazzei, F. *et al* (2013). Role of MUTYH in human cancer. *Mutat Res* 743-744, pp. 33-43.

Mazzoni, SM. *et al* (2015). An AXIN2 Mutant Allele Associated With Predisposition to Colorectal Neoplasia Has Context-Dependent Effects on AXIN2 Protein Function. *Neoplasia* 17(5), pp. 463-472.

Mazzoni, SM. and Fearon, ER. (2014). AXIN1 and AXIN2 variants in gastrointestinal cancers. *Cancer Lett* 355(1), pp. 1-8.

Mignone, F. *et al* (2002). Untranslated regions of mRNAs. *Genome Biol* 3(3), p. reviews0004.1.

Milbury, CA. *et al*. (2011). COLD-PCR: improving the sensitivity of molecular diagnostics assays. *Expert Rev Mol Diagn* 11(2), pp. 159-169.

Mishra, N. and Hall, J. (2012). Identification of patients at risk for hereditary colorectal cancer. *Clin Colon Rectal Surg* 25(2), pp. 67-82.

Mishra, R. *et al.* (2007). Glycogen synthase kinase-3beta induces neuronal cell death via direct phosphorylation of mixed lineage kinase 3. *The Journal of biological chemistry* 282 pp. 30393–405.

Miyaki, M. *et al* (2005). Germline mutations of the MYH gene in Japanese patients with multiple colorectal adenomas. *Mut Res* 5;578(1-2), pp. 430-3.

Miyaki, M. and Kuroki, T. (2003). Role of Smad4 (DPC4) inactivation in human cancer. *Biochem Biophys Res Commun* 306(4), pp. 799-804.

Miyoshi, Y. *et al* (1992a). Germ-line mutations of the APC gene in 53 familial adenomatous polyposis patients. *Proc Natl Acad Sci U S A* 89(10), pp. 4452-4456.

Miyoshi, Y. (1992b). Somatic mutations of the APC gene in colorectal tumors: mutation cluster region in the APC gene. *Hum Mol Genet* 1(4), pp. 229-33.

Mongin, C. *et al* (2012). Unexplained polyposis: a challenge for geneticists, pathologists and gastroenterologists. *Clin Genet* 81(1), pp.38-46.

Nakamura, Y. *et al* (2016). Tissue- and stage-specific Wnt target gene expression is controlled suüsequent to β-catenin recruitment to cis-regulatory modules. *Development* 143(11), pp. 1914-1915

Nielsen, M. *et al* (2007). Germline mutations in APC and MUTYH are responsible for the majority of families with attenuated familial adenomatous polyposis. *Clin Genet* 71(5), pp. 427-433.

Nieminen, TT. *et al* (2011). BMPR1A mutations in hereditary nonpolyposis colorectal cancer without mismatch repair deficiency. *Gastroenterology* 141(1), pp. e23-26.

Nishisho, I. *et al* (1991). Mutations of chromosome 5q21 genes in FAP and colorectal cancer patients. *Science* 253(5020), pp. 665-669.

Nohmi, T. *et al* (2005). Modulation of oxidative mutagenesis and carcinogenesis by polymorphic forms of human DNA repair enzymes. *Mutat Res* 591(1-2), pp. 60-73.

O'Brien, MJ. *et al* (1990). The National Polyp Study. Patient and polyp characteristics associated with high-grade dysplasia in colorectal adenomas. *Gastroenterology* 98(2), pp. 371-379.

Office for National Statistics (2017). Deaths registered in England and Wales: 2016. Statistical bulletin, release date 19th July, 2017. Available at https://www.ons.gov.uk/peoplepopulationandcommunity/birthsdeathsandmarriages/deaths/bulletins/deathsregistrationsummarytables/2016 (accessed 04/12/2017).

Olschwang, S. *et al* (1998). Somatically acquired genetic alterations in flat colorectal neoplasias. *Int J Cancer* 77, pp. 366-369.

Olschwang, S. *et al* (1993). Germ-line mutations in the first 14 exons of the adenomatous polyposis coli (APC) gene. *Am J Hum Genet* 52(2), pp. 273-279.

Omundsen, M. and Lam, FF. (2012). The other colonic polyposis syndromes. *ANZ J Surg* 82(10), pp. 675-681.

Palles, C. *et al* (2013). Germline mutations affecting the proofreading domains of POLE and POLD1 predispose to colorectal adenomas and carcinomas. *Nat Genet* 45(2), pp. 136-144.

Paspatis, GA. *et al* (2001). Prevalence of polyps and diverticulosis of the large bowel in the Cretan population. An autopsy study. *Int J Colorectal Dis* 16(4), pp. 257-261.

Pavicic, W. *et al* (2014). Promoter-specific alterations of APC are a rare cause for mutation-negative familial adenomatous polyposis. *Genes Chromosomes Cancer* 53(10), pp. 857-864.

Pendergrass, CJ. *et al* (2008). Occurrence of colorectal adenomas in younger adults: an epidemiologic necropsy study. *Clin Gastroenterol Hepatol* 6(9), pp. 1011-1015.

Peutz, J (1921). Over een zeer merkwaardige, gecombineerde familiaire polyposis van de slijmliezen van den tractus intestinalis met die van de neuskeelholte en gepaard met eigenaardige pigmentaties van huid-en slijmvliezen. *Ned Maandschr v Gen* (10), pp. 134-146.

Pharoah, PD. *et al* (2001). Incidence of gastric cancer and breast cancer in CDH1 (E-cadherin) mutation carriers from hereditary diffuse gastric cancer families. *Gastroenterology* 121(6), pp. 1348-1353.

Poulsen, ML. and Bisgaard, ML. (2008). MUTYH Associated Polyposis (MAP). *Curr Genomics* 9(6), pp. 420-435.

Powell, SM. *et al* (1993). Molecular diagnosis of familial adenomatous polyposis. *N Engl J Med* 329(27), pp. 1982-1987.

Powell, SM. *et al* (1992). APC mutations occur early during tumorigenesis. *Nature* 359, pp. 235-237.

Rahman, N. (2014). Realizing the promise of cancer predisposition genes. *Nature* 505(7483), pp. 302-308.

Richards, FM. *et al* (1999). Germline E-cadherin gene (CDH1) mutations predispose to familial gastric cancer and colorectal cancer. *Hum Mol Genet* 8(4), pp. 607-610.

Rivera, B. *et al* (2013). A novel AXIN2 germline variant associated with attenuated FAP without signs of oligodontia or ectodermal dysplasia. *Eur J Hum Genet* 22(3), pp. 423-426.

Rohlin, A. *et al* (2011). Inactivation of promoter 1B of APC causes partial gene silencing: evidence for a significant role of the promoter in regulation and causative of familial adenomatous polyposis. *Oncogene* 30(50), pp. 4977-4989.

Rohlin, A. *et al* (2009). Parallel sequencing used in detection of mosaic mutations: comparison with four diagnostic DNA screening techniques. *Hum Mutat* 30(6), pp. 1012-1020.

Rosty, C. *et al* (2012). Phenotype and polyp landscape in serrated polyposis syndrome: a series of 100 patients from genetics clinics. *Am J Surg Pathol* 36(6), pp. 876-882.

Rosty, C. *et al* (2013a). Serrated polyps of the large intestine: current understanding of diagnosis, pathogenesis, and clinical management. *J Gastroenterol* 48(3), pp. 287-302.

Rosty, C. *et al* (2013b). Multiplicity and molecular heterogeneity of colorectal carcinomas in individuals with serrated polyposis. *Am J Surg Pathol* 37(3), pp. 434-442.

Rowan, AJ. *et al* (2000). APC mutations in sporadic colorectal tumors: A mutational "hotspot" and interdependence of the "two hits". *Proc Natl Acad Sci U S A* 97(7), pp. 3352-3357.

Rozen, P. and Baratz, M. (1982). Familial juvenile colonic polyposis with associated colon cancer. *Cancer* 49(7), pp. 1500-1503.

Rozen, S. and Skaletsky, H. (2000). Primer3 on the WWW for general users and for biologist programmers. *Methods Mol Biol* 132, pp. 365-386.

Ruggieri, V. *et al* (2013). Loss of MUTYH function in human cells leads to accumulation of oxidative damage and genetic instability. *Oncogene* 32(38), pp. 4500-4508.

Rundle, AG. *et al* (2008). Colonoscopic screening in average-risk individuals ages 40 to 49 vs 50 to 59 years. *Gastroenterology* 134(5), pp. 1311-1315.

Russell, AM. *et al* (2006). Prevalence of MYH germline mutations in Swiss APC mutation-negative polyposis patients. *Int J Cancer* 118(8), pp. 1937-1940.

Ryland, GL. *et al* (2015). Loss of heterozygosity: what is it good for? *BMC Med Genomics* 8, p. 45.

Sampson, JR. *et al* (2003). Autosomal recessive colorectal adenomatous polyposis due to inherited mutations of MYH. *Lancet* 362(9377), pp. 39-41.

Sarasqueta, AF. *et al* (2013). Integral analysis of p53 and its value as prognostic factor in sporadic colon cancer. *BMC Cancer* 13, p. 277.

Sarles, JC. *et al* (1987). Mixed Familial Polyposis Syndromes *Int J Colorectal Dis*. 2(2), pp. 96-99.

Sasieni, PD. *et al* (2011). What is the lifetime risk of developing cancer? The effect of adjusting for multiple primaries. *Br J Cancer* 105(3), pp. 460-465.

Scheurle, D. *et al* (2000). Cancer gene discovery using digital differential display. *Cancer Res* 60(15), pp. 4037-4043.

Schiemann, AH. and Stowell, KM. (2016). Comparison of pathogenicity prediction tools on missense variants in RYR1 and CACNA1S associated with malignant hyperthermia. *Br J Anaesth* 117(1), pp. 124-128.

Scott, RJ. *et al* (2001). Familial adenomatous polyposis: more evidence for disease diversity and genetic heterogeneity *Gut* 48, pp. 508-514.

Seger, R. and Krebs, EG. (1995). The MAPK signalling cascade. *FASEB J* 9(9), pp. 726-735.

Shen, J. *et al* (2011). Allelic imbalance in BRCA1 and BRCA2 gene expression and familial ovarian cancer. *Cancer Epidemiol Biomarkers Prev* 20(1), pp. 50-56.

Shields, JM. *et al* (2000). Understanding Ras: 'it ain't over 'til it's over'. *Trends Cell Biol* 10(4), pp. 147-154.

Shinya, H. and Wolff, WI. (1979). Morphology, anatomic distribution and cancer potential of colonic polyps. *Ann Surg* 190(6), pp. 679-683.

Short E. *et al* (2015). Inherited predisposition to colorectal cancer: towards a more complete picture. *J Med Gen* 52(12), pp. 791-796.

Sieber, OM. *et al* (2003). Multiple colorectal adenomas, classic adenomatous polyposis, and germ-line mutations in MYH. *N Engl J Med* 348(9), pp. 791-799.

Signori, E. *et al* (2001). A somatic mutation in the 5'UTR of BRCA1 gene in sporadic breast cancer causes down-modulation of translation efficiency. *Oncogene* 20(33), pp. 4596-4600.

Snover, DC. *et al* (2010). Serrated polyps of the colon and rectum and serrated ("hyperplastic") polyposis. In: *World Health Organisation Classification of Tumours of the Digestive System, 4th Edition.* Ed. Bosman, ST. *et al* Pub. Springer-Verlag, Berlin.

Soon, MS. *et al* (2005). Screening colonoscopy in Chinese and Western patients: a comparative study. *Am J Gastroenterol* 100(12), pp. 2749-2755.

South, CD. *et al* (2009). Immunohistochemistry staining for the mismatch repair proteins in the clinical care of patients with colorectal cancer. *Genet Med* 11(11), pp. 812-817.

Spier, I. *et al* (2016). Low-level APC mutational mosaicism is the underlying cause in a substantial fraction of unexplained colorectal adenomatous polyposis cases. *J Med Gen* 53(3), pp.172-179.

Spier, I. *et al* (2012). Deep intronic APC mutations explain a substantial proportion of patients with familial or early-onset adenomatous polyposis. *Hum Mutat* 33(7), pp. 1045-1050.

Strul, H. *et al* (2006). The prevalence rate and anatomic location of colorectal adenoma and cancer detected by colonoscopy in average-risk individuals aged 40-80 years. *Am J Gastroenterol* 101(2), pp. 255-262.

Su, LK. *et al* (2000). Genomic rearrangements of the APC tumor-suppressor gene in familial adenomatous polyposis. *Hum Genet* 106(1), pp. 101-107.

Suchy, J. *et al* (2010). CHEK2 mutations and HNPCC-related colorectal cancer. *Int J Cancer* 126(12), pp. 3005-3009.

Sulová, M. *et al* (2007). Mutation analysis of the MYH gene in unrelated Czech APC mutation-negative polyposis patients. *Eur J Cancer* 43(10), pp. 1617-1621.

Swan, MK. *et al* (2009). Structural basis of high-fidelity DNA synthesis by yeast DNA polymerase delta. *Nat Struct Mol Biol* 16(9), pp. 979-986.

Sweet, K. *et al* (2005). Molecular classification of patients with unexplained hamartomatous and hyperplastic polyposis. *JAMA,* 16:294(19), pp. 2465-2473.

Taupin, D. *et al* (2015). A deleterious RNF43 germline mutation in a severely affected serrated polyposis kindred. *Hum Gen Var* 16(2), pp. 15013.

Terry, MB. *et al* (2002). Risk factors for advanced colorectal adenomas: a pooled analysis. *Cancer Epidemiol Biomarkers Prev* 11(7), pp. 622-629.

Thompson, BA. *et al* (2013). Calibration of multiple in silico tools for predicting pathogenicity of mismatch repair gene missense substitutions. *Hum Mutat* 34(1), pp. 255-265.

Thylur, RP. *et al.* (2011). Mixed lineage kinase 3 modulates β-catenin signaling in cancer cells. *The Journal of biological chemistry* 286, pp. 37470–82.

Toyota, M. *et al* (1999). CpG island methylator phenotype in colorectal cancer. *Proc Natl Acad Sci U S A* 96(15), pp. 8681-8686.

Tuohy, TM. and Burt, RW. (2008). Somatic mosaicism: a cause for unexplained cases of FAP? *Gut* 57(1), pp. 10-12.

Tuohy, TM. *et al* (2010). Large intron 14 rearrangement in APC results in splice defect and attenuated FAP. *Hum Genet* 127(3), pp. 359-369.

Tutlewska, K. *et al* (2013). Germline deletions in the EPCAM gene as a cause of Lynch syndrome - literature review. *Hered Cancer Clin Pract* 11(1), pp. 9.

Valizadeh, A. *et al*. (1997). Expression of E-cadherin-associated molecules (alpha-, beta-, and gamma-catenins and p120) in colorectal polyps. *Am J Pathol* 150(6), pp. 1977-1984.

Valle, L. *et al* (2014). New InSiGHTs into POLE and POLD1 germline mutations in familial colorectal cancer and polyposis. *Hum Mol Genet* 23(13), pp. 3506-3512.

Van der Velde, KJ. *et al* (2015). Evaluation of CADD Scores in Curated Mismatch Repair Gene Variants Yields a Model for Clinical Validation and Prioritization. *Hum Mutat* 36(7), pp. 712-719.

Van Wyk, R. *et al* (1999). Multiple APC mutations in sporadic flat colorectal adenomas. *Eur J Hum Genet* 7, pp. 928-932.

Vaughn, CP. *et al* (2010). Clinical analysis of PMS2: mutation detection and avoidance of pseudogenes. *Hum Mutat* 31(5), pp. 588-593.

Venesio, T. *et al* (2012) MUTYH-associated polyposis (MAP), the syndrome implicating base excision repair in inherited predisposition to colorectal tumors. *Front Oncol* 2, p. 83.

Venesio, T. *et al* (2004). High frequency of MYH gene mutations in a subset of patients with familial adenomatous polyposis. *Gastroenterology* 126(7), pp. 1681-1685.

Vogelstein, B. *et al* (1988). Genetic alterations during colorectal-tumor development. *N Engl J Med* 319(9), pp. 525-532.

Vogelstein, B. and Kinzler, KW (2004). Cancer genes and the pathways they control. *Nat Med* 10(8), pp. 789-799.

Wagner, JR. *et al* (2010). Computational analysis of whole-genome differential allelic expression data in human. *PLoS Comput Biol* 6(7), pp. e1000849.

Walsh, MD. *et al* (2012). Immunohistochemical testing of conventional adenomas for loss of expression of mismatch repair proteins in Lynch syndrome mutation carriers: a case series from the Australasian site of the colon cancer family registry. *Mod Pathol* 25(5), pp. 722-730.

Walsh, T. *et al* (2006). Spectrum of mutations in BRCA1, BRCA2, CHEK2, and TP53 in families at high risk of breast cancer. *JAMA* 295(12), pp. 1379-1388.

Walther, A. *et al* (2009). Genetic prognostic and predictive markers in colorectal cancer. *Nat Rev Cancer* 9, pp. 489-499.

Wang, L. *et al* (2004). MYH mutations in patients with attenuated and classic polyposis and with young-onset colorectal cancer without polyps. *Gastroenterology* 127(1), pp. 9-16.

Wen, J. and Brogna, S. (2008). Nonsense-mediated mRNA decay. *Biochem Soc Trans* 36(Pt 3), pp. 514-516.

Weren, RD. *et al* (2015). A germline homozygous mutation in the base-excision repair gene NTHL1 causes adenomatous polyposis and colorectal cancer. *Nat Genet* 47(6), pp. 668-671.

Westerman, AM. *et al* (1999). Peutz-Jeghers syndrome: 78-year follow-up of the original family. *Lancet* 353(9160), pp. 1211-1215.

Wilding, JL. *et al* (2010). Replication error deficient and proficient colorectal cancer gene expression differences caused by 3'UTR polyT sequence deletions. *Proc Natl Acad Sci USA* 107(49), pp. 21058-21063.

Williams, AR. *et al* (1982). Polyps and cancer of the large bowel: a necropsy study in Liverpool. *Gut* 23(10), pp. 835-842.

Williams, GT. et al (1980). Metaplastic polyps and polyposis of the colorectum. *Histopathology* 4(2), pp. 155-170.

Win, AK. *et al* (2014). Risk of Colorectal Cancer for Carriers of Mutations in MUTYH, With and Without a Family History of Cancer. *Gastroenterology* 146(5), pp. 1208-1211.

Wong, SC. *et al* (2004). Prognostic and diagnostic significance of beta-catenin nuclear immunostaining in colorectal cancer. *Clin Cancer Res* 10(4), pp. 1401-1408.

Woodford-Richens, KL. *et al* (2001). SMAD4 mutations in colorectal cancer probably occur before chromosomal instability, but after divergence of the microsatellite instability pathway. *Proc Natl Acad Sci U S A* 98(17), pp. 9719-9723.

Wu, ZQ. *et al* (2012). Canonical Wnt suppressor, Axin2, promotes colon carcinoma oncogenic activity. *Proc Natl Acad Sci U S A* 109(28), pp. 11312-11317.

Yanus GA. *et al* (2018). Spectrum of APC and MUTYH germ-line mutations in Russian patients with colorectal malignancies. *Clin Genet* doi:10.1111/cge.13228 (Epub ahead of print).

Yamaguchi, K. *et al* (2016). Reduced expression of APC-1B but not APC-1A by the deletion of promoter 1B is responsible for familial adenomatous polyposis. *Sci Rep* 6, p. 26011.

Yamaji, Y. *et al* (2004). Incidence and recurrence rates of colorectal adenomas estimated by annually repeated colonoscopies on asymptomatic Japanese. *Gut* 53(4), pp. 568-572.

Yamane, L. *et al* (2014). Serrated pathway in colorectal carcinogenesis. *World J Gastroenterol* 20(10), pp. 2634-2640.

Yan, H. *et al* (2002a). Allelic variation in human gene expression *Science* 297, pp. 1943.

Yan, H. *et al* (2002b). Small changes in expression affect predisposition to tumorigenesis. *Nature Genetics* 30, pp. 25-26.

Yan, HHN. *et al* (2017). RNF43 germline and somatic mutation in serrated neoplasia pathway and its association with BRAF mutation. *Gut* 66(9), pp. 1645-1656.

Yeoman, A. *et al* (2007). Hyperplastic polyposis in the New Zealand population: a condition associated with increased colorectal cancer risk and European ancestry. *N Z Med J* 120(1266), p. U2827.

Yu, HMI. *et al* (1995). The Role of AXIN2 in Calvarial Morphogenesis and Craniosyntosis. *Dev* 132(8), 1995-2005.

Zhou, XL. *et al* (2005). Germline mutations in the MYH gene in Swedish familial and sporadic colorectal cancer. *Genet Test* 9(2), pp. 147-151.