Author/s:
Lung, Mei Sim

Title:
Identification of candidate genes predisposing to familial colorectal cancer by germline whole exome sequencing

Date:
2019

Persistent Link:
https://hdl.handle.net/11343/239184

# Identification of candidate genes predisposing to familial colorectal cancer by germline whole exome sequencing

Mei Sim Lung

ORCID ID: 0000-0001-9849-144X

Submitted in total fulfilment of the requirements of the degree of Doctor of Philosophy

September, 2019

Faculty of Medicine, Dentistry and Health Sciences

The Sir Peter MacCallum Department of Oncology

The University of Melbourne

# Abstract

Colorectal cancer (CRC) is the third most common cause of cancer-related death worldwide. Approximately 5% of CRC is attributable to an inherited high-penetrance pathogenic variant in a known CRC-predisposing gene, often associated with a family history of CRC, a young onset of CRC and/or colonic polyposis. Identification of an inherited genetic variant predisposing to CRC in an individual with CRC enables predictive testing in blood relatives, and implementation of risk reduction strategies for individuals with the causative variant within the family. In individuals with familial or young onset CRCs/colonic polyposis in whom a causative germline variant in a known CRC-predisposing gene has not been identified, it is hypothesised that novel CRC-predisposing genes harbouring rare high risk germline variants remain to be discovered.

Germline whole exome sequencing (WES) offered a novel and comprehensive approach to identifying potential pathogenic variants in candidate CRC-predisposing genes. This thesis used WES of germline DNA in individuals with unexplained CRC and a high *a priori* risk of an inherited predisposition to CRC, to prioritise candidate CRC-predisposing genes that were subsequently re-sequenced in a validation cohort of young-onset CRC cases.

A discovery cohort was recruited comprising 54 individuals who met one or more of the following criteria: CRC and a family history of CRC (n=29), a diagnosis of CRC before 40 years of age (n=20), or a diagnosis of 20 or more bowel polyps (n=10). No known germline predisposition to CRC had been identified in these individuals through clinical testing. Rare (minor allele frequency of <0.5%) loss-of-function (nonsense, frameshift or essential splice site) variants or missense variants predicted to be deleterious using a Combined Annotation Dependent Depletion C Score of >10 were prioritised for analysis. Three individuals with pathogenic variants in known CRC-predisposing genes were identified and excluded from further analyses.

In the remaining 51 individuals, four separate analysis approaches were used to prioritise candidate CRC-predisposing genes containing prioritised variants: 1) genes with variants in multiple families, 2) genes with a shared variant in multiple family members, 3) genes within known CRC pathways, and 4) genes intolerant of loss-of-function variants. Forty-five candidate CRC-predisposing genes were prioritised using these analyses. Segregation analysis of potentially pathogenic variants within a family (n=1) and loss-of-heterozygosity studies in tumour tissue (n=4) were performed but hampered by lack of availability of samples.

Re-sequencing of the 45 candidate CRC-predisposing genes in a validation cohort of 333 individuals with unexplained young onset CRC identified seven genes (*WNT10A, NEDD4, MS4A15, TNFRSF10B, LRRK2, INPP5E, UFL1)* with a loss-of-function variant in 1 or more individuals. None of the seven genes harboured LoF variants in >1% of the validation cohort, indicating that a novel gene likely accounts for very few individuals with unexplained CRC and a high *a priori* risk of an inherited predisposition to CRC, pointing to a very high degree of genetic heterogeneity in such cases. Future studies will require very large

numbers of cases and controls to confirm the association of these genes with CRC predisposition.

# Declaration

This is to certify that:

(i) the thesis comprises only my original work towards the PhD except where indicated in the preface

(ii) due acknowledgement has been made in the text to all other material used

(iii) the thesis is fewer than 100 000 words in length, exclusive of tables, maps, bibliographies and appendices

Mei Sim Lung

# Preface

This thesis involved several collaborations, and due acknowledgement is given throughout this thesis.

I would like to thank the familial cancer clinics across Australia who provided clinical samples, and the families who agreed to participate in the study.

I am indebted to the Research Computing Services at the Peter MacCallum Cancer Centre, Dr. Maria Doyle, Dr Jason Li and Richard Lupat, for their support and implementation of the variant calling pipeline, and to Simone Rowley for her support and guidance on molecular techniques.

I would also like to thank my supervisors Professor Ian Campbell and Associate Professor Alison Trainer for their support and advice on this thesis.

Massive Parallel Sequencing of DNA was performed through BGI (Shenzhen, Guangdong, China), the Australian Genome Research Facility and the Molecular Core Facility at Peter MacCallum Cancer Centre.

Local sequencing datasets used during variant filtering and variant analysis were provided by Dr. Sally Hunter and Li Na of the Campbell Lab at Peter MacCallum Cancer Centre.

I am grateful to Dr. Catherine Mitchell from the Pathology Department at Peter MacCallum Cancer Centre for assessing CRC samples for tumour purity.

A multi-author manuscript included in Chapter 4 has been submitted for publication to BMC Cancer on 19/09/2019.

Title: Germline whole exome sequencing of a family with appendiceal mucinous tumours presenting with pseudomyxoma peritonei

Authors: **Mei Sim Lung**, Catherine A. Mitchell, Maria A. Doyle, Andrew C. Lynch, Kylie L. Gorringe, David D.L. Bowtell, Australian Ovarian Cancer Study Group, Ian G. Campbell, Alison H. Trainer

I am the first author on this multi-author publication and contributed >50% of the work, including planning of the study, whole genome amplification, Sanger sequencing, tumour microdissection, analysis and interpretation of whole exome sequencing data and tumour data, drafting and revision of the manuscript. The contributions of the co-authors is described in detail in the manuscript, and include histopathology expertise, bioinformatics support, sample provision and advisory comments. All co-authors have signed the co-author forms submitted with this thesis.

The work described in this thesis was funded by an Australian National Health and Medical Research Council and Victorian Cancer Council grant (APP1028374). I was funded by an Australia Postgraduate Award from the University of Melbourne, Australia.

# Acknowledgements

I would firstly like to thank the Giver of Life for sustaining me through this journey and pursuit of knowledge.

I wish to express heartfelt gratitude to my co-supervisors Professor Ian Campbell and Associate Professor Alison Trainer for all your advice, supervision, compassion and kindness in this research apprenticeship. Your impartation of genomics knowledge and mentoring in academic rigour have forever changed the way I approach the practice of oncology.

I gratefully acknowledge the mentorship given by Dr. Gisela Mir Arnau, Dr. Sarah-Jane Dawson and Dr. Sherene Loi at various times.

I would like to thank Simone Rowley, Dr. Ella Thompson, Dr. Sally Hunter, Dr. Georgina Ryland, Dr. Dane Cheasley, Associate Professor Kylie Gorringe and Dr. David Goode for your support in the Campbell Lab. Thank you also to members of the student committee and Associate Professor Caroline Owen for your support during my studentship.

This research would not have been possible without the clinicians, genetic counsellors, staff and patients of Familial Cancer Centres across Australia; thank you for your collaboration and participation in this study that sought to address an unmet need in the clinic.

I owe so much to my family, in particular my father Dr. Jien Kien Lung who passed away in the first year of this PhD, my mum Chui Fui Chow for her indefatiguable and sacrificial love and support, and my brother Shyan for his love and support.

To members of the Campbell Lab past and present, and to the research community at Peter MacCallum Cancer Centre, each day of this research student life was made richer by your friendship and intellectual discourse. Special thanks to Simone, Ella, Dane, Kenny, David, Deepak, Tanjina, Kylie, Na, Sally Davis, Amanda, Abhi, Lisa, Carolina, Hugo, Colin, and members of the Peter Mac Soccer Team.

Thank you also to my church family, Neville and Aurellia Bion, Dirk and Nevan Visser, Zora, Tze-Min, Jacqueline, Sharon, Daniel, Carol and Mark Penny. I would not have survived without your prayers, understanding and love.

# Table of Contents

# List of Tables

# List of Figures

# Abbreviations

| Abbreviation | Definition |
|---|---|
| A | adenine |
| AMTs | appendiceal mucinous tumours |
| *APC* | adenomatous polyposis coli |
| BAM | binary sequence alignment |
| BDT | Big Dye Terminator |
| BMP | bone morphogenetic protein |
| **bp** | base pair |
| BWA | Burrows-Wheeler Algorithm |
| °C | degrees Celsius |
| C | cytosine |
| CI | confidence interval |
| CRC | colorectal cancer |
| DNA | deoxyribonucleic acid |
| dNTPs | deoxynucleotide solution mix |
| ESP | Exome Sequencing Project |
| EtOH | ethanol |
| ExAC | Exome Aggregation Consortium |
| FAP | familial adenomatous polyposis |
| FCC | familial cancer clinic |
| FCCTX | Familial Colorectal Cancer Type X |
| FDR | first degree relative |
| FFPE | formalin-fixed paraffin-embedded |
| G | guanine |
| GATK | Genome Analysis Toolkit |
| GO | Gene Ontology |
| GWAS | genome-wide association studies |
| H&E | haemotoxylin and eosin |
| HMPS | Hereditary Mixed Polyposis Syndrome |
| HNPCC | Hereditary Non Polyposis Colorectal Cancer |
| HREC | Human Research and Ethics Committee |
| IHC | immunohistochemistry |
| indel | insertion/deletion |
| JPS | juvenile polyposis syndrome |
| kb | kilobase |
| kConFab | Kathleen Cuningham Foundation Consortium for Research into Familial Aspects of Breast Cancer |
| L | litre |
| LoF | loss-of-function |
| LOH | loss of heterozygosity |
| µg | microgram |
| µL | microlitre |
| mL | millilitre |
| MLPA | multiplex ligation-dependent probe amplification |
| MMR | mismatch repair |
| NHLBI GO ESP | National Heart, Lung and Blood Institute Greater Opportunity Exome Sequencing Project |
| ng | nanogram |
| NZ | New Zealand |

| Abbreviation | Definition |
|---|---|
| PCR | polymerase chain reaction |
| PJS | Peutz-Jeghers Syndrome |
| pLI | probability of being loss-of-function intolerant |
| PMCC | Peter MacCallum Cancer Centre, Melbourne, Australia |
| PMP | pseudomyxoma peritonei |
| PolyPhen | polymorphism phenotyping |
| PPVs | potentially pathogenic variants |
| PTT | protein truncation test |
| QUAL | base calling Phred quality score |
| rpm | revolutions per minute |
| **RVIS** | Residual Variation Intolerance Score |
| SDR | second degree relative |
| SIFT | Sorting Intolerant from Tolerant |
| SNP | single nucleotide polymorphism |
| SNV | single nucleotide variant |
| T | thymidine |
| TCGA | The Cancer Genome Atlas |
| USA | United States of America |
| VCB | Victorian Cancer Biobank |
| VQSR | variant quality score recalibration |
| WES | whole exome sequencing |

# Chapter 1. Introduction

The purpose of this thesis was to identify novel genes predisposing to colorectal cancer (CRC) in individuals suspected of having a hereditary predisposition, but in whom no established CRC-predisposing gene mutation had been found. This study explores the use of whole exome sequencing (WES) of germline DNA in a cohort of such individuals in order to identify novel genetic variants predisposing to CRC. In this chapter, the known genetic aetiologies of CRC are reviewed, together with the methods used for their discovery. The thesis aims and overview are then presented.

## 1.1. The Importance of Hereditary Colorectal Cancer

World Health Organisation data show that Australia and New Zealand combined have the highest estimated incidence rates of CRC in the world, at 41.7 per 100,000 men and 32.1 per 100,000 women in 2018, seven to eight-fold higher than South-Central Asia that has the lowest rate (World Health Organisation, 2018). Worldwide, CRC is the third most common cancer in men and second in women, and the third most common cause of cancer-related mortality (Jemal et al., 2011).

CRC arises from colonic epithelium via an accumulation of genetic and/or epigenetic aberrations. The most well-known pathway of progression to CRC is described in the archetypical adenoma-carcinoma model proposed by Fearon and Vogelstein (Fearon and Vogelstein, 1990). In this pathway, loss of function of the tumour suppressor gene adenomatous polyposis coli (*APC)* results in hyperproliferation of the colonic epithelium. This is followed by DNA hypomethylation and inactivation of tumour suppressor genes *DCC* on chromosome 18q and *TP53* on chromosome 17p, together with activating mutations in *KRAS*, resulting in the development of a colonic adenoma followed by transformation to CRC. These CRCs exhibit chromosomal instability (Fodde et al., 2001). A second pathway accounting for a smaller sub-group of CRCs is initiated by loss of function of a mismatch repair (MMR) gene (*MLH1*, *MSH2*, *MSH6* or *PMS2),* which results in multiple frameshift mutations at microsatellites (Aaltonen et al., 1993). These tumours are often diploid but microsatellite unstable. A third pathway, termed the CpG island methylator phenotype pathway, is characterised by hypermethylation across the genome of the CRC, and is associated with methylation and silencing of genes such as *MLH1* (Toyota et al., 1999*)*. CRCs associated with the CpG island methylator phenotype pathway often originate in serrated lesions (Jass, 2007a).

While the majority of CRC cases occur sporadically, approximately 5% are currently attributed to inheritance of a rare, high-penetrance pathogenic variant that predisposes to CRC and other cancers, while up to 10% are attributed to inheritance of common low penetrance CRC risk alleles (Carvajal Carmona and Tomlinson, 2016, Al-Tassan et al., 2015). The hereditary CRCs arising from rare, high penetrance pathogenic variants are important entities to recognise for several reasons.

Firstly, while the risk of developing CRC to age 85 years in Australia is 1 in 11 for men and 1 in 16 for women (Australian Institute of Health and Welfare, 2017), for individuals with hereditary CRC, this risk is significantly increased and usually associated with a younger age of

onset. An example from the highest end of the penetrance spectrum is individuals with a germline *APC* mutation, who have a 100% risk of developing CRC by age 40 years (Bisgaard et al., 1994). Identifying individuals and their families who may have inherited a rare, high-penetrance CRC-predisposing gene mutation is vital, as risk-reducing strategies such as screening colonoscopy have been associated with improved survival (Järvinen* et al., 2000, Renkonen-Sinisalo et al., 2000a, Heiskanen et al., 2000b, De Jong et al., 2006). Identification of the causative pathogenic variant within a family allows predictive testing to be performed, and intensive surveillance to be offered only to family members who carry the family mutation. Predictive testing also benefits related individuals who do not carry the family mutation and hence do not require intensive surveillance. The health system can also benefit as a result from decreased resource utilisation.

Secondly, the elucidation of the genes underlying hereditary CRC has had broader applicability in sporadic CRC, both in terms of understanding tumorigenesis and selecting treatment. The *APC* gene that is mutated in the germline of patients with familial adenomatous polyposis, was later found to be mutated in the majority of sporadic precursor adenomas and CRCs (Mori et al., 1992, Powell et al., 1992). *MLH1*, a mismatch repair gene that is mutated in Lynch syndrome, was subsequently found to be methylated in sporadic CRCs with microsatellite instability (Herman et al., 1998), and defines an important subgroup of CRC that responds to immunotherapy (Le et al., 2015). Identification of new pathogenic variants in hereditary CRC may help to prioritise somatic variants identified in sequencing studies of CRC. Inherited pathogenic variants are likely to be cancer-initiating, hence a corresponding somatic mutation is likely to be a driver mutation, helping to distinguish it from passenger mutations. The genes that are known to predispose to hereditary CRC when mutated are discussed in detail in the following section.

## 1.2. Inherited Rare, High-Penetrance, Gene Mutations Predisposing to CRC

Historically, hereditary CRC was divided into 2 groups based on the presence or absence of polyposis. The first group consists of CRC with associated colonic polyposis. It is sub-divided based on polyp histology: adenomatous polyposis due to mutations in the *APC* or *MUTYH* genes (Sieber et al., 2003); hamartomatous polyposis due to *SMAD4, BMPR1A* or *STK11* mutations; and hereditary mixed polyposis syndrome due to duplications of *GREM1* (Jaeger et al., 2012). The genetic aetiology of a fourth subgroup, serrated polyposis, is yet to be determined. The second group, which gives rise to hereditary non-polyposis CRC (later renamed Lynch syndrome when the genetic aetiology was determined), is caused by mutations or epigenetic silencing of the MMR genes *MLH1, MSH2, MSH6* and *PMS2.*

The clinical syndromes arising from these gene mutations were recognised prior to the identification of the underlying genetic cause, and indeed were used to select appropriate patients for gene discovery. Subsequently, it was found that some of these genes can give rise to CRC with or without colonic polyposis. *MUTYH*, which was first identified in patients with adenomatous polyposis, can give rise to early onset CRC in the absence of polyposis (Farrington et al., 2005). MMR gene mutations were first identified in the heterozygous state in the germline of patients with hereditary non-polyposis colorectal cancer. When mutated in a homozygous state in the germline, MMR gene mutations result in polyposis (Jasperson et al., 2011). The extra-colonic cancers associated with each gene can also differ, and is described in more detail below.

Because of this phenotypic heterogeneity, identification of the underlying pathogenic variant is important for both accurate classification and screening, as the clinical course and cancer spectrum may differ. In this section, the clinical features of the polyposis syndromes (familial adenomatous polyposis, hamartomatous polyposis, hereditary mixed polyposis syndrome, serrated polyposis) and Lynch syndrome are described, followed by the method of identification of the associated colorectal cancer predisposing genes and important genotype-phenotype correlations.

### 1.2.1. Familial Adenomatous Polyposis (FAP)

#### *1.2.1.1. Clinical Phenotype*

Classic FAP is characterised by the development of hundreds to thousands of colorectal adenomatous polyps by the second decade of life. By the fourth decade, CRC develops in virtually 100% of cases if the polyps are left *in situ* (Bisgaard et al., 1994). CRC arises in the left side of the colon in 70-80% of cases (Björk, 1999). FAP accounts for less than 1% of all CRC cases.

FAP is also associated with extra-colonic cancers and benign tumours (Groen et al., 2008), including duodenal and gastric adenomas and cancer, medulloblastoma, rarely papillary thyroid cancer and childhood hepatoblastoma, non-malignant desmoid tumours, congenital hypertrophy of the retinal pigment epithelium that is often bilateral (Trainer, 2009), lipomas and adrenal tumours.

An attenuated form of FAP is observed, with less than 100 adenomatous polyps in the colon, and a delayed diagnosis of CRC by 15 years (Knudsen et al., 2010).

#### *1.2.1.2. Discovery of the APC Gene*

The chromosome region containing the *APC* gene was identified from a case report in 1986, of a 42-year-old patient with an interstitial deletion of 5q15-22 presenting with synchronous colon cancers, more than 100 adenomatous colonic polyps, a desmoid tumour, mental retardation and a constellation of other developmental abnormalities (Herrera et al., 1986). The locus was further refined to 5q21-q22 by studying five polymorphic chromosome 5 DNA probes in 124 individuals from 13 families with classic FAP (Bodmer et al., 1987). At the same time, a similar finding was independently reported using linkage analysis using 16 DNA markers in 4 large FAP families, 2 of whom had Gardner syndrome (Leppert et al., 1987). Further mapping was performed by linkage analysis using 6 new polymorphic DNA markers in 6 FAP families, 3 of whom had Gardner syndrome (Nakamura et al., 1988). The chromosome region of interest was narrowed down further by the finding of two deletions spanning 100 kb and 260 kb respectively in two unrelated FAP patients, with the 100 kb deletion being nested in the larger deletion (Joslyn et al., 1991). Three genes were found in this region: *SRP19*, *DP1* (deleted in polyposis 1, also known as *REEP5*) and *DP2.5*. Single strand conformation polymorphism analysis of these 3 genes followed by sequencing of unique conformers in 61 unrelated FAP patients and 12 controls revealed 4 inactivating mutations in DP2.5 (two nonsense and two frameshift) in 4 FAP patients, and no variants unique to patients in the other 2 genes. *DP2.5* was thus identified as the gene for FAP and was renamed *APC*.

#### *1.2.1.3. Genotype-Phenotype Correlations*

Germline *APC* mutations give rise to an autosomal dominant disorder, although an estimated 25% of germline *APC* mutations occur *de novo* (Bisgaard et al., 1994). *APC* mutations are protein-truncating in 95% of cases (Fearnhead et al., 2001). A common pathogenic missense mutation is found in at least 5% of the Ashkenazi Jewish population (Rozen et al., 1999b, Liang et al., 2013). The position of the mutation within the APC gene influences the clinical phenotype. For example, congenital hypertrophy of the retinal pigment epithelium is strongly associated with mutations occurring between codons 311 and 1444, while mutations in codon 1309 result in an earlier age of onset of CRC (mean age 35 years) (Bertario et al., 2003). Mutations beyond codon 1444 have a 12-fold increased risk of desmoid tumours (Bertario et al., 2001). More severe polyposis (>5000 polyps) develops with mutations between codons 1250 and 1464 (Nagase et al., 1992). Attenuated FAP was initially associated with mutations in the 5' end, exon 9 and 3' end of *APC*, however mutations have subsequently been found throughout the gene in these patients (Knudsen et al., 2010).

#### *1.2.1.4. Discovery of the MUTYH Gene*

*MUTYH* was found to be a CRC-predisposing gene through the study of British family N (Al-Tassan et al., 2002). Family N consisted of 3 siblings with polyposis in a sibship of 7, with unaffected parents. Two of the affected siblings had approximately 50 colonic adenomas on colectomy at ages 59 and 55 years respectively, while the third sibling died of CRC at age 46 years. No inheritedpathogenic variant was identified in *APC*. Sequencing of *APC* in 11 tumours from Family N was performed and revealed that six of eight tumours harboured biallelic somatic inactivation of *APC*. Eighty-three percent (15 of 18) of the somatic inactivating events

were a distinct class of G:C to T:A transversions. This was significantly elevated when compared to somatic *APC* mutations from sporadic CRCs and adenomas, and somatic *APC* mutations from FAP tumours, in which the rate of G:C to T:A transversions was approximately 10%.

Increased G:C to T:A transversions secondary to oxidised guanine mispairing with adenine had been seen in *Escherichia coli*, with mutation of the *MutY* gene (Moriya and Grollman, 1993), and also in *Saccharomyces cereviciae* with inactivation of the *OGG1* gene (Thomas et al., 1997). The coding regions of the human homologues *MUTYH, OGG1* and *MTH* (that prevents incorporation of oxidised guanine into DNA (Michaels and Miller, 1992)) were sequenced in the germline of an individual from family N. This revealed two missense variants in exons 7 and 13 of *MUTYH* that encoded for non-conservative amino acid changes p.Tyr165Cys and p.Gly382Asp respectively. No potentially pathogenic variants were identified in *OGG1* or *MTH*.

The two other affected members of family N were also found to be compound heterozygotes for the *MUTYH* mutations, while unaffected family members were either heterozygous for only one of the mutations or did not carry either *MUTYH* mutation. Screening for the two *MUTYH* mutations in 100 British controls identified one heterozygous case for each mutation.

Biallelic inherited pathogenic variants in *MUTYH* were subsequently found in 7 out of 21 unrelated British patients with more than 100 adenomas (Jones et al., 2002). In another study, Sieber et al screened the *MUTYH, OGG1* and *MTH1* genes in two groups of patients: 1)patients with 3 to 100 synchronous or metachronous colon adenomas, and 2)patients with >100 colon adenomas in whom no germline *APC* mutation had been identified, as well as in 107 unaffected British controls (Sieber et al., 2003). Six out of 152 patients who presented with adenomas were found to have biallelic germline *MUTYH* mutations, while another 6 had monoallelic *MUTYH* mutations. Three of the patients with biallelic mutations were compound heterozygotes, while the other 3 were presumed homozygous. While one patient with compound heterozygous mutations had both the p.Tyr165Cys and p.Gly382Asp mutations, the other two patients were found to have novel 1 bp deletions in addition to either the p.Tyr165Cys or p.Gly382Asp mutation. No biallelic *MUTYH* mutations were found in British controls.

#### *1.2.1.5. Genotype-Phenotype Correlations*

The FAP-like phenotype arising from biallelic *MUTYH* mutations, henceforth referred to as *MUTYH*-Associated Polyposis (MAP) is an autosomal recessive condition. The p.Tyr165Cys and p.Gly382Asp mutations are European founder mutations accounting for 73% of *MUTYH* mutations in Western populations, albeit with a probable reporting bias (Cheadle and Sampson, 2007).

Homozygous p.Tyr179Cys mutations give rise to a more severe phenotype, with a higher risk of CRC (56 fold vs. 19 fold) and a younger mean age of diagnosis (49.5 years vs. 57.9 years) compared to homozygous p.Gly396Asp mutations (Lubbe et al., 2009). The severity of the phenotype of the p.Gly396Asp/p.Tyr179Cys compound heterozygote lies in between the two homozygous groups with a median age of onset of CRC of 52.5 years (Lubbe et al., 2009). Nonsense, small insertion/deletions (indels) and splice site variants have also been reported (Cheadle and Sampson, 2007).

A mixture of adenomatous and serrated polyps is seen in 50% of patients with biallelic germline *MUTYH* mutations (Guarinos et al., 2014). Serrated polyps are associated with the p.Gly396Asp pathogenic variant (Guarinos et al., 2014). Serrated polyps from patients with biallelic germline *MUTYH* mutations tend to harbour somatic *KRAS* p.Gly12Cys mutations instead of BRAFV600E mutations (Guarinos et al., 2014), presumably because G:C to T:A transversions would give rise to a cysteine in place of glutamine.

Extra-colonic manifestations include duodenal and gastric fundic polyposis (Terdiman, 2009). Cases of ovarian cancer, endometrial cancer, sebaceous gland tumours and testicular cancer have also been reported in patients with biallelic germline *MUTYH* mutations (Vogt et al., 2009, Guarinos et al., 2014). As mentioned previously, patients with biallelic *MUTYH* mutations may develop young onset CRC in the absence of polyposis.

There are conflicting reports on the risk of CRC in patients with heterozygous germline *MUTYH* mutations. An older case control study that included a meta-analysis of other case-control studies showed no increased risk (Balaguer et al., 2007) (odds ratio 1.11, 95% confidence interval (CI) 0.9-1.37), while several newer studies that included a wider range of *MUTYH* mutations suggest an increased risk (Win et al., 2014) (hazard ratio 2.46 for men (95% CI 1.54-3.93) and 2.67 for women (95% CI 1.67-4.26)), (Cleary et al., 2009) (adjusted odds ratio 1.48 (95% CI 1.02-2.16)), (Jones et al., 2009a) (standardised incidence ratio 2.12 (95% CI 1.30-3.28)).

### 1.2.2. Hamartomatous Polyposis

Hamartomas are benign tumours, consisting of cells from the tissue of origin growing in a disorganised way. Nevertheless, the hamartomatous polyposis syndromes juvenile polyposis syndrome (JPS) and Peutz-Jeghers syndrome (PJS) are associated with an increased risk of CRC and various malignancies.

#### *1.2.2.1. Peutz-Jeghers Syndrome (PJS)*

##### 1.2.2.1.1. Clinical Phenotype

PJS is an autosomal dominant disorder with the hallmark, mucocutaneous pigmentation, seen in 95% of cases. In PJS hamartomatous polyps are found throughout the gastro-intestinal tract with patients presenting in childhood with bowel obstruction, intussusception or bleeding. The lifetime risk of CRC was found to be 39%, although the rarity of the condition may lead to ascertainment and publication bias. Associated extra-colonic cancers include breast, pancreatic, gastric, gynaecological and testicular (Jasperson et al., 2010, Hearle et al., 2006).

##### 1.2.2.1.2. Discovery of the *STK11* Gene

*STK11* (also known as *LKB1*) was identified in 1998 (Hemminki et al., 1998). The location of this gene on chromosome 19p was identified by studying the DNA from multiple hamartomatous polyps from an individual with PJS using comparative genomic hybridisation. This approach identified a common loss of chromosome 19p. These somatic deletions were consistently found on the chromosome from the unaffected parent, suggesting that the causative gene could be a tumour suppressor following Knudsen's two-hit hypothesis with retention of the mutated gene from the affected parent. The region of interest was further refined by meiotic

recombination studies in PJS patients. Transcripts in this region were sought, and the *LKB1* gene, which encodes a serine/threonine kinase, was sequenced in 12 PJS patients. Nine of 12 familial PJS patients were found to carry a heterozygous nonsense or frameshift mutation in *LKB1*. Studies in another 4 PJS individuals also identified nonsense or acceptor splice site mutations (Jenne et al., 1998).

### *1.2.2.2. Juvenile Polyposis Syndrome (JPS)*

#### 1.2.2.2.1. Clinical Phenotype

A clinical diagnosis of JPS requires 1 of 3 criteria: (1) more than 5 juvenile polyps in the colorectum; (2) juvenile polyps throughout the gastrointestinal tract; (3) any number of juvenile polyps with a family history of juvenile polyposis. Patients usually develop per rectal bleeding, abdominal obstruction or anaemia in childhood. Risk of CRC is between 9-68%, and associated cancers include pancreatic, stomach and duodenal (Pyatt et al., 2006). Cleft palate and congenital heart disease has also been identified in up to 20% of JPS patients (Desai et al., 1998).

#### 1.2.2.2.2. Discovery of the *SMAD4* Gene

A tightly linked region between two markers on 18q21.1 was identified by linkage analysis in a large juvenile polyposis family (Howe et al., 1998). Two genes were located in this region: *DCC* and *SMAD4*. Exon sequencing of *SMAD4* revealed a 4 bp (base pair) deletion in exon 9 in one of the affected individuals. This mutation was subsequently identified in all 13 affected family members and in none of 7 spouses and 242 unrelated individuals. Sequencing of all *SMAD4* exons in a further eight unrelated juvenile polyposis patients identified 2 individuals with a 4 bp deletion in exon 9, one individual with a 2 bp deletion in exon 8, and another individual with a 1 bp insertion in exon 5. No *SMAD4* mutations were identified in the remaining 4 affected individuals. All identified mutations were predicted to cause protein truncation.

#### 1.2.2.2.3. Discovery of the *BMPR1A* Gene

Linkage analysis was undertaken in the four juvenile polyposis families in whom no mutations in *SMAD4* were identified. This approach identified a candidate region at 10q22-23 (Howe et al., 2001b) that harboured the gene *BMPR1A,* which was known to interact with *SMAD4*. Linkage analysis using markers closer to *BMPR1A* showed a highly significant LOD score of >4. Sequencing of the four families identified two different nonsense variants, a 4 bp deletion and a 1 bp deletion. None of these mutations were identified in 250 normal controls. Both familial mutations segregated with disease within the families. JPS is an autosomal dominant disorder.

#### 1.2.2.2.4. Genotype-phenotype correlations

In a study of 54 juvenile polyposis patients, in which 9 were found to have *SMAD4* mutations and 13 were found to have *BMPR1A* mutations, patients with *SMAD4* mutations were more likely to have a family history of gastric polyposis compared to those with *BMPR1A* mutations (Sayed et al., 2002). *SMAD4* pathogenic variants are also associated with clinical hereditary haemorrhagic telangiectasia in approximately 20% of cases (Aretz et al., 2007), although the genotypic /phenotypic correlation is unclear. Therefore all patients with a *SMAD4* mutation require screening for JPS and hereditary haemorrhagic telangiectasia that can present with

mucocutaneous telangiectasia, lung, cerebral or liver arteriovenous malformations or recurrent epistaxis.

### 1.2.3. Hereditary Mixed Polyposis Syndrome (HMPS)

#### *1.2.3.1. Clinical Phenotype*

The term HMPS was first used to describe the mixed histopathologies often seen within single polyps from a large polyposis family at St. Mark's Hospital in England (Whitelaw et al., 1997). This Jewish family had been followed for 40 years over 4 generations, with 42 individuals developing either CRC or colonic polyps. While the defining feature of this syndrome is the development of mixed histopathologies within a polyp, specifically atypical juvenile polyps with mixed adenomatous and/or hyperplastic components, the majority of affected family members developed tubular adenomas. The polyposis was inherited in an autosomal dominant manner. In contrast to FAP, less than 15 colonic polyps were seen on initial colonoscopy in affected individuals, and polyps occurred throughout the colon. In contrast to Lynch syndrome, CRCs in these individuals did not show preponderance to a particular site in the colon, and were microsatellite stable. The median age of developing CRC was 40 years. There were no extra-colonic manifestations noted in affected individuals, in particular no ocular or cutaneous features of FAP, or peri-oral freckling associated with PJS. Upper gastro-intestinal tract assessment was not reported in the original study.

#### *1.2.3.2. Discovery of the GREM1 Duplication*

The genetic variant giving rise to HMPS was found by systematic interrogation of the chromosome region on chromosome 15q13.3 identified by linkage analysis of two affected Jewish families (Jaeger et al., 2012). The coding regions of the three genes within the region of interest, *GREM1, SCG5* and *FMN1* were sequenced in 3 affected individuals and 2 unaffected relatives who did not share the haplotype at this chromosome region, but no gene mutation was found that segregated with disease. The observation that additional patients with the same clinical phenotype were all of Jewish origin led to the postulation that a founder mutation was responsible for HMPS. Copy number analysis of the region was then performed by a custom oligonucleotide array, and this uncovered a 40 kb simple tandem duplication separated by a 30 bp insertion of a unique sequence that was mapped to chr15:30,752,231-30,792,051 using Human Genome Build 36.
PCR primers were designed to span the duplication boundary, producing a unique 190 bp product. A further 40 cases and 50 controls from 6 HMPS families were screened. The 190 bp product was only found in cases. One hundred and eighty-eight unselected Ashkenazi Jewish controls were screened, and the 190 bp product was not found. In order to exclude other pathogenic variants within the haplotype region identified by linkage analysis, the entire haplotype region was sequenced and no other pathogenic variant was identified.

The duplication commenced at intron 2 of *SCG5* and ended upstream of the CpG island of *GREM1.* The functional significance of this duplication was elicited by mRNA expression studies, which found normal levels of *SCG5* transcripts, but increased levels of *GREM1* transcripts in normal colonic crypts of HMPS patients compared to that from unaffected

controls. Moreover, GREM1 expression was limited to the crypt base in unaffected individuals but extended up the sides of the crypts in HMPS patients. Further functional studies using chromatin immunoprecipitation and chromatin conformation capture revealed transcription factor-binding sites within the duplication that interacted with the *GREM1* promoter to increase gene expression.

### 1.2.4. Serrated Polyposis

#### *1.2.4.1. Clinical Phenotype*

Serrated polyps consist of hyperplastic polyps, sessile serrated adenomas and traditional serrated polyps. The World Health Organisation classification of serrated polyposis requires the following: (1) more than 5 serrated polyps proximal to the sigmoid colon with 2 of these being more than 10mm; or (2) any number of serrated polyps proximal to the sigmoid colon in an individual who has a first-degree relative with serrated polyposis; or (3) more than 20 serrated polyps of any size, but distributed throughout the colon (Snover et al., 2010). Serrated polyposis is phenotypically heterogeneous, with some patients having predominantly left-sided polyps, others having large right-sided polyps, and a third group having serrated polyps throughout the colon (Kalady et al., 2011, Jass, 2007b). This raises the possibility of underlying locus heterogeneity or genetic/environmental modifiers giving rise to these three subgroups. In a retrospective study of 100 probands referred to familial cancer clinics with serrated polyposis, the prevalence of CRC was 31%, with an average age at diagnosis of 48 years (Win et al., 2012). CRC and pancreatic cancer were associated with serrated polyposis (Win et al., 2012).

#### *1.2.4.2. Gene Discovery*

The genetic aetiology of serrated polyposis is still being elucidated. *RNF43* nonsense mutations have been reported in 2 families with serrated polyposis (Yan et al., 2016, Taupin et al., 2015), and 2 out of 20 unrelated individuals with serrated polyposis (Gala et al., 2014a). However, WES of 72 individuals with serrated polyposis did not identify any further protein truncating *RNF43* variants, and high-resolution melt analysis of the 2 previously identified *RNF43* nonsense variants did not identify any further variants in 221 serrated polyposis individuals (Buchanan et al., 2016). A homozygous germline *MUTYH* mutation was found in 1 of 38 patients with serrated polyposis (Chow et al., 2006), but this was not validated in a larger cohort of 65 individuals with serrated polyposis, nor were pathogenic variants in other colonic polyposis predisposing genes found (Clendenning et al., 2013). Further linkage studies have identified a region on chromosome 2q32.2-2q33.3 as a candidate region, although no gene mutations in the region of interest were identified (Roberts et al., 2011).

### 1.2.5. Lynch Syndrome

Recognition of this clinical syndrome began with the report of a family designated Family G in whom 18 of 48 individuals from 3 successive generations had developed cancer, many in their 40s (Warthin, 1913). Ten women had uterine cancer, 7 individuals had gastric cancer and at least 2 individuals had intestinal cancer. Fifty-three years later, Henry Lynch (after whom this

syndrome is named) and colleagues described two more families (Families N and M) from Nebraska and Michigan in the United States of America, who were noted to have a similar preponderance to uterine, stomach and intestinal cancers in 4 and 3 successive generations respectively (Lynch et al., 1966). The syndrome was called Hereditary Non-Polyposis Colorectal Cancer (HNPCC), due to the lack of colonic polyposis (distinguishing this syndrome from familial adenomatous polyposis, caused by *APC* mutations.

The Amsterdam I criteria (Vasen et al., 1991) were developed to identify suitable families for further research into the genetic aetiology of this syndrome. The criteria consisted of:

1. at least 3 family members affected with CRC, of whom one must be a first-degree relative of the other two. CRCs have to be verified with histology, and FAP must be ruled out.

2. CRC must be seen in at least two successive generations

3. one person must have developed CRC before age 50 years

A proportion of HNPCC cases were subsequently found to be caused by an autosomal dominant MMR gene defect. These cases were renamed as Lynch Syndrome cases, to differentiate these families from families meeting Amsterdam I criteria but in whom a genetic predisposition had not been determined. The latter have subsequently been designated 'Familial CRC Type X (FCCTX)' (Lindor et al., 2005).

#### *1.2.5.1. Clinical Phenotype*

The median age of developing CRC is 45 years in individuals with Lynch syndrome (Lynch et al., 2009), compared to 71 years in the Australian population (Australian Institute of Health and Welfare, 2017). Several histopathologic features are characteristic of CRCs in Lynch syndrome, including poor differentiation, presence of tumour inflitrating lymphocytes, signet ring cell or mucinous histology and location proximal to the splenic flexure (Mecklin et al., 1986, Lynch et al., 1991a) .

Extra-colonic tumours associated with Lynch syndrome include endometrial cancer, ovarian cancer, gastric cancer, small bowel cancer, transitional cell carcinoma of the renal pelvis, ureters or bladder, glioblastomas (known as Turcot variant) and skin cancers such as sebaceous gland tumours and keratoacanthomas (known as Muir-Torre variant) (Watson et al., 2008a, Lynch et al., 2009 ).

Constitutive Mismatch Repair Deficiency refers to the inheritance of biallelic MMR gene mutations, often resulting in haematological malignancies and gliomas in childhood, as well as Lynch syndrome tumours and café au-lait spots on the skin (Felton et al., 2007). As alluded to earlier, in contrast to monoallelic pathogenic variants in the MMR genes, biallelic pathogenic variants in MMR genes can give rise to colonic adenomatous polyposis.

#### *1.2.5.2. Discovery of the MMR Genes*

The discovery of the mismatch repair genes as colorectal cancer-predisposing genes began with genome-wide linkage analyses of two large HNPCC families from Canada and New Zealand (Peltomaki et al., 1993). The chromosomal region of interest in these families was narrowed down to chromosome 2p15-16. In concurrent research, DNA from CRCs in HNPCC patients was noted to have replication errors in the form of alterations in the length of

dinucleotide and trinucleotide repeat sequences (Aaltonen et al., 1993, Lynch et al., 1991b).This was termed microsatellite instability (MSI). Similar replication errors had been seen in *Sacchomyces cereviciae* (Strand et al., 1993) harbouring mutations in the genes *msh2, mlh1* and *pms1* that are involved in a mismatch repair system first characterised in bacteria (Modrich, 1991).

The human homologue to *msh2, MSH2* was cloned and found to map to chromosome 2p21 (Fishel et al., 1993), close to the region identified in the original linkage analysis (Peltomaki et al., 1993). Partial sequencing of the *MSH2* gene in 2 HNPCC families identified a heterozygous splice acceptor site mutation that segregated with disease (Fishel et al., 1993). A similar approach of linkage analysis in HNPCC families (Lindblom et al., 1993) led to the identification of a locus at 3p21 and cloning of the human homologue of *mlh1.* Analysis of the *MLH1* gene in HNPCC families identified a missense mutation in affected individuals (Bronner et al., 1994).

The roles of other mismatch repair genes in HNPCC were then investigated by identifying the human homologues of additional *Sacchomyces cereviciae* and *Escherichia coli* mismatch repair genes (Peltomaki, 2005). Seven more human homologues were identified, however only pathogenic variants in *MSH6* (Akiyama et al., 1997*,* Miyaki et al., 1997) were found to segregate with disease in HNPCC families. A *PMS2* pathogenic variant was identified in a HNPCC patient (Nicolaides et al., 1994), and in 3 patients with Turcot variant HNPCC (Peltomaki, 2005). The accurate detection of germline *PMS2* mutations is hampered by the presence of pseudogenes (Nakagawa et al., 2004), requiring the use of long-range PCR to increase specificity (Clendenning et al., 2006).

MLH1, MSH2, MSH6 and PMS2 function within the MMR pathway, maintaining genomic integrity by identifying and removing base-pairing errors or insertion deletion loops arising during DNA replication (Kunkel and Erie, 2005). MSH2 and MSH6 form a heteroduplex, known as MutSα, that recognises and initiates repair (Palombo et al., 1995, Acharya et al., 1996). The MSH2-MSH6 heteroduplex also recruits the MLH1-PMS2 heteroduplex, known as MutLα (Acharya et al., 1996), to mismatched sites (Habraken et al., 1998, Blackwell et al., 2001). The MLH1-PMS2 heteroduplex is activated by proliferating cell nuclear antigen to incise the nascent strand, allowing removal of the mismatched bases (Kadyrov et al., 2006, Kadyrov et al., 2007). MMR proteins are also involved in other cellular processes including somatic hypermutation, class switch recombination in immunoglobulin genes and double strand break repair (Martin and Scharff, 2002, Cascalho et al., 1998, Martomo et al., 2004).

Pre-genetic screening of tumours with immunohistochemistry (IHC) stains for MLH1, MSH2, MSH6 and PMS2 is now used to identify individuals with Lynch syndrome. As MSH6 relies on MSH2 for protein expression while PMS2 relies on MLH1, loss of MHL1 results in corresponding loss of PMS2 on IHC, while loss of MSH2 results in corresponding loss of MSH6 (Lynch et al., 2009). Loss of MSH6 or PMS2 results in isolated loss of the corresponding protein in the tumour on IHC. Hence, loss of protein expression on IHC directs germline sequencing efforts to the appropriate gene. Missense mutations that do not result in nonsense-mediated decay of the corresponding protein can result in a false negative result with IHC. Genotyping for microsatellite instability is also used to identify Lynch syndrome tumours, and can identify Lynch syndrome cases missed by IHC due to expression of non-functional protein.

### *1.2.5.3. Discovery of deletions in the 3' end of the EPCAM gene leading to epigenetic silencing of MSH2*

A unique genetic mechanism leading to Lynch syndrome that did not involve overt mutations of the MMR genes was identified by Ligtenberg *et al* in 2008 and Kovacs *et al* in 2009. A Dutch family with MSI-high CRC and *MSH2* loss on tumour IHC, was found not to have a pathogenic variant in *MSH2.* However, a reduced signal on multiplex ligation-dependent probe amplification (MLPA) analysis of *MSH6* for exon 9 of *TACSTD1* (also known as EPCAM*)*, 16 kb upstream of *MSH2* (Ligtenberg et al., 2009) was noted. Using more focused MLPA analysis and long-range PCR, a deletion involving the last two exons of *EPCAM*, but not involving the promoter of *MSH2*, was identified. The same deletion was found in a further 3 unrelated Dutch families with MSH2-deficient MSI-high CRC without germline *MSH2* mutations, suggesting this may be a founder mutation. The deletion was not found in 145 Dutch CRC patients being investigated for *MSH6* mutations, or in 97 healthy Dutch controls. The deletion segregated with disease in 2 families, and in CRCs from 1 of the families, somatic *MSH2* mutations were found. A different germline deletion again involving the 3' end of *EPCAM* but without involvement of the *MSH2* promoter was seen in 2 unrelated Chinese families with CRCs containing promoter methylation of *MSH2*. The CRCs from the affected Dutch families were also found to have methylation of the *MSH2* promoter. Similarly MLPA of genomic DNA in 27 Hungarian HNPCC families that were negative for *MLH1* and *MSH2* mutations showed a decrease in signal strength corresponding to deletions in the 3' region of the *EPCAM* gene in 5 families. The deletions were found to segregate with CRC in these families, and were not found in 45 individuals from 21 control families. The tumours from these families were found to be microsatellite unstable and had loss of *MSH2* on IHC. Deletions of the 3' end of the *EPCAM* gene including its polyadenylation signal result in transcriptional readthrough across the CpG island of the promoter of *MSH2* downstream (Ligtenberg et al., 2009). This results in promoter methylation in tissues that express *EPCAM* such as colonic epithelium.

### *1.2.5.4. Genotype-phenotype correlations*

Lynch syndrome is an autosomal dominant disorder, while Constitutive Mismatch Repair Deficiency is a recessive condition. A study of 23 Lynch syndrome cases identified from 1066 unselected CRCs found that mutations in *MSH2* were most common (n=13), followed by *MLH1* (n=5), *MSH6* (n=3) and *PMS2* (n=2) (Hampel et al., 2005). Ninety-one percent (21/23) of the detected pathogenic variants in this study resulted in protein truncation. As genotyping for microsatellite instability was the initial test used for identification of Lynch syndrome cases, it is unlikely to have biased mutation detection towards truncating variants. Truncating mutations similarly accounted for 94% of 333 Australian Lynch syndrome families ascertained on clinical criteria or microsatellite testing/MMR protein IHC over a 14 year period (Sjursen et al., 2016).

Missense mutations are more common in *MLH1* than *MSH2* (Woods et al., 2007). Founder mutations have been described, such as the *MSH2* c.942+3A>T splice site mutation resulting in the deletion of exon 5 in affected families from Newfoundland, Canada (Froggatt et al., 1999).This pathogenic variant also occurs de novo worldwide (Desai et al., 2000) and may account for up to 10% of Lynch syndrome cases (Hampel et al., 2005).

The risks of developing colorectal cancer and extra-colonic cancers vary for each of the Lynch syndrome genes, as demonstrated in a prospective observational study (Møller et al., 2018) and in retrospective studies (Bonadona et al., 2011, Engel et al., 2012, Watson et al., 2008b). Retrospective studies are limited to risk estimates due to the need to adjust for ascertainment bias. Table 1.1 summarises the Lynch syndrome-associated cancer risks to age 75 years, according to MMR gene subtype, from the Prospective Lynch Syndrome Database (Møller et al., 2018). These prospectively determined risks await validation in an independent prospective study.

**Table 1-1.Lynch syndrome-associated cancer risks to age 75 years from the Prospective Lynch Syndrome Database (Møller et al., 2018)**

| Mutated MMR Gene | Risk of cancer type to age 75 (%) | | | | | | |
|---|---|---|---|---|---|---|---|
| | Colorectal | Endometrial | Ovarian | Gastric | Duodenal | Urinary tract | Brain |
| *MLH1* | 58.5 | 42.7 | 10.1 | 7.1 | 6.5 | 8.7 | 1.0 |
| *MSH2* | 60.7 | 56.7 | 16.9 | 7.7 | 2.0 | 25.9 | 5.3 |
| *MSH6* | 18.8 | 46.2 | 13.1 | 5.3 | 0 | 11.2 | 1.4 |
| *PMS2* | 0 | 26.4 | 0 | 0 | 0 | 0 | 0 |

Patients with *EPCAM* deletions have a much lower rate of endometrial cancer compared with patients with *MSH2* mutations, likely due to the differential expression of *EPCAM* in colorectal and endometrial tissues (Kempers et al., 2011). Patients with germline *EPCAM* mutations who developed endometrial cancer were noted to have deletions that extended closest to the *MSH2* promoter (less than 2.5 kb).

## 1.3. A Proportion of Each Clinical Phenotype Has No Known Predisposing Gene

In each of the clinical phenotypes described above, there are a proportion of cases in whom a CRC-predisposing gene mutation has not been identified (Table 1.2). In the case of FAP, full sequencing and Southern blot analysis of the *APC* gene, and sequencing of exons 7 and 13 of the *MUTYH* gene, together with full sequencing of the *MUTYH* gene in cases where only one of the two most common *MUTYH* mutations was found, was performed in 1457 individuals with classic FAP and 3252 individuals with attenuated FAP (Grover et al., 2012). In individuals with 10-999 adenomatous polyps, no mutations were found in 37-91% of cases, while in individuals with ≥1000 adenomatous polyps, no mutation was identified in 18% of cases. In PJS, *STK11* mutations are identified in up to 80% of cases (Hearle et al., 2006), while *SMAD4* or *BMPR1A* mutations are seen in 40-60% of JPS cases (Howe et al., 2004, Aretz et al., 2007). In serrated polyposis, only 1 of 40 individuals was found to have biallelic germline *MUTYH* mutations, suggesting that these are rare in serrated polyposis (Guarinos et al., 2014). The genetic factors predisposing to the majority of serrated polyposis cases have not yet been determined. In the case of families meeting Amsterdam-I criteria, mutations in *MLH1* and *MSH2* were only identified in 45% of cases (Wijnen et al., 1998). When *MSH6* and *PMS2* mutations are taken into account, together with improved mutation detection, it is still estimated that 40-50% of families meeting Amsterdam-I criteria do not have an inherited pathogenic variant in a MMR gene (Lindor, 2009).

**Table 1-2. Proportion of hereditary CRC phenotypes explained by the known CRC-predisposing genes**

| Clinical Phenotype | Gene | Percentage of cases with a Gene Mutation | Percentage of unexplained cases |
|---|---|---|---|
| Attenuated adenomatous polyposis (10-19 polyps) (Grover et al., 2012) | *APC* | 5% | 91% |
| | *MUTYH* | 4% | |
| Attenuated adenomatous polyposis (20-99 polyps) (Grover et al., 2012) | *APC* | 10% | 83% |
| | *MUTYH* | 7% | |
| Classic familial adenomatous polyposis (100-999 polyps) (Grover et al., 2012) | *APC* | 56% | 37% |
| | *MUTYH* | 7% | |
| Classic familial adenomatous polyposis (≥1000 polyps) (Grover et al., 2012) | *APC* | 80% | 18% |
| | *MUTYH* | 2% | |
| Peutz-Jegher Syndrome (Hearle et al., 2006) | *STK11* | ≤80% | At least 20% |
| Juvenile Polyposis (Howe et al., 2004, Aretz et al., 2007) | *SMAD4* | 40-60% | At least 40% |
| | *BMPR1A* | | |
| Hereditary Mixed Polyposis | *GREM1* | N/A | N/A |
| Serrated polyposis(Guarinos et al., 2014) | *MUTYH* | 2.5% | N/A |
| | *RNF43* | N/A | |
| Families meeting Amsterdam-I criteria (Wijnen et al., 1998) | *MLH1* | 45% | Up to 55% |
| | *MSH2* | | |
| | *MSH6* | N/A | |
| | *PMS2* | N/A | |
| | *EPCAM* | N/A | |

N/A, not available.

## 1.4. Common Low Penetrance Variants Associated with CRC

The contribution of common germline variants to CRC predisposition has been investigated through genome-wide association studies (GWAS). GWAS are case-control studies that aim to identify 'risk' alleles that are more associated with disease using the 'common disease-common allele' approach. Oligonucleotide microarray chips are used to assay the allele frequencies of thousands of single nucleotide polymorphisms (SNPs) distributed throughout the human genome in cases and controls. As this results in multiple testing of large numbers of SNPs, the commonly accepted threshold of genome-wide significance for such studies is usually set at a stringent value of $5 \times 10^{-8}$ (Altshuler et al., 2005, Pe'er et al., 2008) in order to reduce false positives.

GWAS have identified over 40 alleles associated with increased CRC risk, with odds ratios ranging from 1.06-1.35 (Table 1.3) (Schumacher et al., 2015, Whiffin et al., 2014, Wang et al., 2014, Zhang et al., 2014, Peters et al., 2013, Jia et al., 2013, Peters et al., 2012, Dunlop et al., 2012, Tomlinson et al., 2011, Cui et al., 2011, Hutter et al., 2010, Houlston et al., 2010, Tenesa et al., 2008, Jaeger et al., 2008, Houlston et al., 2008, Tomlinson et al., 2008, Zanke et al., 2007, Tomlinson et al., 2007, Broderick et al., 2007, Haiman et al., 2007). Although individually each allele is associated with a small individual increase in risk, in aggregate, these SNPs account for an estimated 10% of the familial CRC risk (Al-Tassan et al., 2015, Carvajal Carmona and Tomlinson, 2016).

While a strength of GWAS is its agnostic approach to risk allele identification, one of the subsequent difficulties is elucidating the functional mechanism by which these SNPs predispose to CRC. As most SNPs are in non-coding regions of the genome, their functional significance was initially attributed to impacting on genes that fell within or adjacent to the region of linkage disequilibrium close to the respective SNP. While this has led to an important observation that genes involved in the TGFβ signalling pathway (*SMAD7, GREM1, BMP2, BMP4* and *RHPN2)* are influenced by risk alleles (Tenesa and Dunlop, 2009), in silico annotation of CRC risk SNPS with mRNA expression in blood cells has suggested that the gene closest to a SNP may not necessarily show the tightest correlation and more distant influences are important (Carvajal-Carmona et al., 2011).

**Table 1-3. Risk loci associated with increased CRC risk from GWAS**

| Chromosome Locus | Risk Allele | rsID | Odds Ratio | 95% CI | Smallest P value | Possible Associated Gene(s) | Reference(s) |
|---|---|---|---|---|---|---|---|
| **15q13** | T | rs4779584 | 1.35 | *1.14-1.60* | *4X10-14* | *GREM1* | (Jaeger et al., 2008, Tomlinson et al., 2008, Peters et al., 2012) |
| **6q25.3** | T | rs7758229 | 1.28 | *1.18-1.39* | *8X10-9* | *SLCC22A3* | (Cui et al., 2011) |
| **8q23.3** | C | rs16892766 | 1.27 | 1.19–1.32 | *3.33X10-18* | *EIF3H* | (Tomlinson et al., 2008) |
| **18q21** | A | rs7229639 | 1.22 | *1.15-1.29* | *3X10-11* | *SMAD7* | (Zhang et al., 2014) |
| **18q21** | T | rs4939827 | 1.20 | *1.16-1.24* | *8X10-28* | *SMAD7* | (Broderick et al., 2007, Tenesa et al., 2008) |
| **8q24.21** | G | rs6983267 | 1.19 | *1.15-1.23* | *9X10-26* | *POU5F1P1, MYC* | (Hutter et al., 2010, Tomlinson et al., 2007, Zanke et al., 2007, Tenesa et al., 2008, Peters et al., 2013, Cui et al., 2011, Haiman et al., 2007) |
| **10q25.2** | C | rs12241008 | 1.19 | *1.12-1.26* | *3X10-8* | *VTI1A* | (Wang et al., 2014) |
| **12p13.32** | T | rs3217810 | 1.19 | *1.13-1.25* | *2.16X10-10* | *CCND2* | (Peters et al., 2013, Whiffin et al., 2014) |
| **5q31.1** | A | rs647161 | 1.17 | 1.11-1.22 | *1.22X10-10* | *PITX1* | (Jia et al., 2013) |
| **12p13.32** | T | rs10774214 | 1.17 | 1.11–1.23 | *5X10-10* | *CCND2* | (Jia et al., 2013) |
| **2q32.3** | C | rs11903757 | 1.16 | 1.10–1.22 | 3.71E-08 | *NABP1* | (Peters et al., 2013) |
| **12q24.22** | G | rs73208120 | 1.16 | *1.11-1.23* | *3X10-8* | *NOS1* | (Schumacher et al., 2015) |
| **11q12.2** | G | rs4246215 | 1.15 | *1.12-1.19* | *7.65X10-20* | *FEN1* | (Zhang et al., 2014) |
| **11q12.2** | T | rs174550 | 1.15 | *1.12-1.19* | *1.58X10-19* | *FADS1* | (Zhang et al., 2014) |
| **11q12.2** | A | rs1535 | 1.15 | *1.12-1.19* | *8.21X10-20* | *FADS2* | (Zhang et al., 2014) |
| **19q13.1** | C | rs10411210 | 1.15 | *1.10-1.20* | *5X10-9* | *RHPN2* | (Houlston et al., 2008) |
| **3p22.1** | A | rs35360328 | 1.14 | *1.09-1.19* | *2.4X10-8* | *CTNNB1* | (Schumacher et al., 2015) |
| **10q25.2** | A | rs11196172 | 1.14 | *1.10-1.18* | *1X10-12* | *TCF7L2* | (Zhang et al., 2014) |
| **12p13.31** | T | rs10849432 | 1.14 | *1.09-1.18* | *3X10-10* | *CD9* | (Zhang et al., 2014) |
| **10p14** | G | rs10795668 | 1.12 | 1.10-1.16 | *3X10-13* | *GATA3* | (Tomlinson et al., 2008) |
| **10q24.2** | T | rs1035209 | 1.12 | *1.08-1.16* | *4.54X10-11* | *ABCC2* | (Whiffin et al., 2014) |

**Table 1.3 (continued). Risk loci associated with increased CRC risk from GWAS**

| Chromosome Locus | Risk Allele | rsID | Odds Ratio | 95% CI | Smallest P value | Possible Associated Gene(s) | Reference(s) |
|---|---|---|---|---|---|---|---|
| **20p12.3** | C | rs961253 | 1.12 | 1.08–1.16 | *2.0X10-10* | *BMP2* | (Houlston et al., 2008) |
| **11q23** | C | rs3802842 | 1.11 | *1.08-1.15* | *5.82X10-10* | *COLCA2* | (Tenesa et al., 2008) |
| **14q22.2** | C | rs4444235 | 1.11 | 1.08–1.15 | *8X10-10* | *BMP4* | (Houlston et al., 2008) |
| **6p21** | A | rs1321311 | 1.10 | *1.07-1.13* | *P* = 1.14 × 10**−10** | *CDKN1A* | (Dunlop et al., 2012) |
| **10q22.3** | G | rs704017 | 1.10 | *1.06-1.13* | *2X10-8* | *ZMIZ1, AS1* | (Zhang et al., 2014) |
| **16q22.1** | G | rs9929218 | 1.10 | *1.06-1.12* | *1X10-8* | *CDH1* | (Houlston et al., 2008) |
| **17p13.3** | C | rs12603526 | 1.10 | *1.06-1.14* | *3X10-8* | *NXN* | (Zhang et al., 2014) |
| **20p12.3** | G | rs4813802 | 1.10 | *1.06-1.12* | *7X10-11* | *BMP2* | (Tomlinson et al., 2011, Peters et al., 2012) |
| **20p12.3** | C | rs2423279 | 1.10 | *1.06-1.14* | *7X10-9* | *BMP2* | (Jia et al., 2013) |
| **1q25.3** | A | rs10911251 | 1.09 | *1.06-1.12* | *1.75X10-8* | *LAMC1* | (Peters et al., 2013, Whiffin et al., 2014) |
| **1q41** | G | rs6687758 | 1.09 | *1.06-1.12* | *2.27X10-9* | *DUSP10* | (Houlston et al., 2010) |
| **3p14.1** | G | rs812481 | 1.09 | *1.05-1.11* | *2.5X10-8* | *LRIG1* | (Schumacher et al., 2015) |
| **12q24.12** | C | rs3184504 | 1.09 | *1.06-1.12* | *2X10-8* | *SH2B3* | (Schumacher et al., 2015) |
| **19q13.2** | G | rs2241714 | 1.09 | *1.06-1.12* | *1X10-8* | *TGFB1* | (Zhang et al., 2014) |
| **20q13.1** | G | rs6066825 | 1.09 | *1.06-1.12* | *4X10-9* | *PREX1* | (Schumacher et al., 2015) |
| **11q13.4** | C | rs3824999 | 1.08 | *1.05-1.10* | 3.65 × 10**−10** | *POLD3* | (Dunlop et al., 2012) |
| **14q22.2** | A | rs1957636 | 1.08 | *1.06-1.11* | 1.36X10-9 | *BMP4* | (Tomlinson et al., 2011) |
| **20q13.33** | C | rs4925386 | 1.08 | *1.05-1.10* | 2X10-10 | *LAMA5* | (Houlston et al., 2010, Peters et al., 2012) |
| **Xp22.2** | C | rs5934683 | 1.07 | *1.04-1.10* | 7.30 × 10**−10** | *SHROOM2* | (Dunlop et al., 2012) |
| **1q41** | T | rs6691170 | 1.06 | 1.03–1.09 | *9.6X10-10* | *DUSP10* | (Houlston et al., 2010) |
| **12q13.13** | T | rs7136702 | 1.06 | *1.04-1.08* | *4X10-8* | *LARP4, DIP2B* | (Houlston et al., 2010) |

## 1.5. WES as a novel strategy for discovering additional rare, Mendelian, familial CRC predisposing genes

As >95% of genetic mutations predisposing to Mendelian diseases affect protein-coding regions (Botstein and Risch, 2003), the use of WES to identify such variants in familial CRC is a targeted and cost-effective strategy. WES has been successfully used to identify the genes predisposing to several Mendelian diseases with rare phenotypes (Ng et al., 2010b, Ng et al., 2010a, Shen et al., 2015, Ku et al., 2011).

### 1.5.1. WES in Familial Cancer Predisposition

At the time of commencement of this study, no familial CRC WES studies had been reported, although its application to elucidating the genetic aetiology of FCCTX had been proposed (Ku et al., 2012). The application of WES in familial pancreatic cancer (Jones et al., 2009c, Roberts et al., 2012), familial phaeochromocytoma (Comino-Méndez et al., 2011), Hodgkin's lymphoma (Saarinen et al., 2011) and breast cancer (Thompson et al., 2012, Park et al., 2012) has successfully identified cancer-predisposing genes.

Following the commencement of this study, pathogenic missense variants in the genes *POLE* and *POLD1* were discovered (Palles et al., 2013) through whole genome sequencing and linkage analysis performed in two families with multiple affected individuals across 3 to 5 consecutive generations, who had developed at least 10 colorectal adenomas by age 60 years, further confirming WES as a worthwhile approach to identifying novel CRC-predisposing genes.

### 1.5.2. Significance of the Study

Identification of novel genes predisposing to familial CRC will allow accurate classification of patients according to genetic aetiology. This will likely result in better disease characterisation based on genotype-phenotype correlations, and enable more accurate cancer risk assessments and preventive strategies. From the point of view of the family of the affected individual, family members can be screened for the family-specific CRC-predisposing gene mutation, and if they do not carry the family-specific gene mutation, they can be excluded from more intensive cancer surveillance that is often recommended for individuals with a genetic predisposition to CRC. This has advantages at both the level of the individual, whereby mutation-positive individuals can be managed effectively in terms of risk-reducing screening strategies, and at the level of healthcare economics, wherein individuals who are not at increased CRC risk are not receiving unnecessary medical intervention.

A greater understanding of the genetic aetiology of CRC may also point to novel pathways of CRC initiation, and better understanding of the biological correlations of genetic alterations can help refine treatment strategies for CRC.

### 1.5.3. Hypothesis, Aim and Scope

The hypothesis of this thesis is that a proportion of individuals with CRC or colonic polyposis who have been referred to a Familial Cancer Centre due to a family history of CRC, a young age of onset of CRC (less than 40 years), and in whom no mutation in a known CRC-predisposing

gene has been found, harbour an inherited predisposing pathogenic variant in a novel CRC-predisposing gene.

The aim of this thesis is to identify novel candidate CRC-predisposing genes using WES of germline DNA from an Australian population deemed to be at increased risk of harbouring an inherited pathogenic variant predisposing to CRC. This study will only examine the coding and essential splice site regions of the human genome. Consideration was given to performing whole genome sequencing instead of WES, however due to the paucity of annotation of the non-coding regions of the genome when this study began, it was decided that investigation of genetic factors in the discovery cohort of this study would be limited to the exome. Using WES rather than whole genome sequencing would also allow for a larger number of patients to be interrogated, as whole genome sequencing at the time this study commenced was approximately three times more expensive than WES.

### 1.5.4. Thesis Objectives and Overview

The objectives of this thesis are as follows:

(1) Identify a cohort of patients with an increased likelihood of harbouring a novel germline CRC-predisposing gene mutation, for the purpose of WES of germline DNA
(2) Identify candidate CRC-predisposing genes through analysis of germline variants from WES
(3) Determine if the candidate CRC-predisposing genes are recurrently mutated in a larger cohort of young-onset CRC through targeted gene sequencing.
(4) Assess for loss-of-heterozygosity of the candidate germline variant by performing Sanger sequencing on DNA from corresponding colonic tumour tissue from the cohort identified in objective (1)
(5) Perform segregation analysis of candidate CRC-predisposing germline variants where additional family members are available and agree to participate

Chapter 3 will address objective (1). Chapter 4 will address objective (2), by presenting the analysis pipeline used to prioritise potentially pathogenic variants identified in WES, followed by the analysis strategies employed to identify candidate CRC-predisposing genes containing potentially pathogenic variants. Chapter 5 will present the re-sequencing of candidate CRC-predisposing genes in a validation cohort, loss-of-heterozygosity studies in corresponding tumour tissue and segregation analyses for objectives (3), (4) and (5). An overview of the results chapters is presented in Figure 1.1. Chapter 6 will summarise the conclusions from these results chapters and discuss future directions.

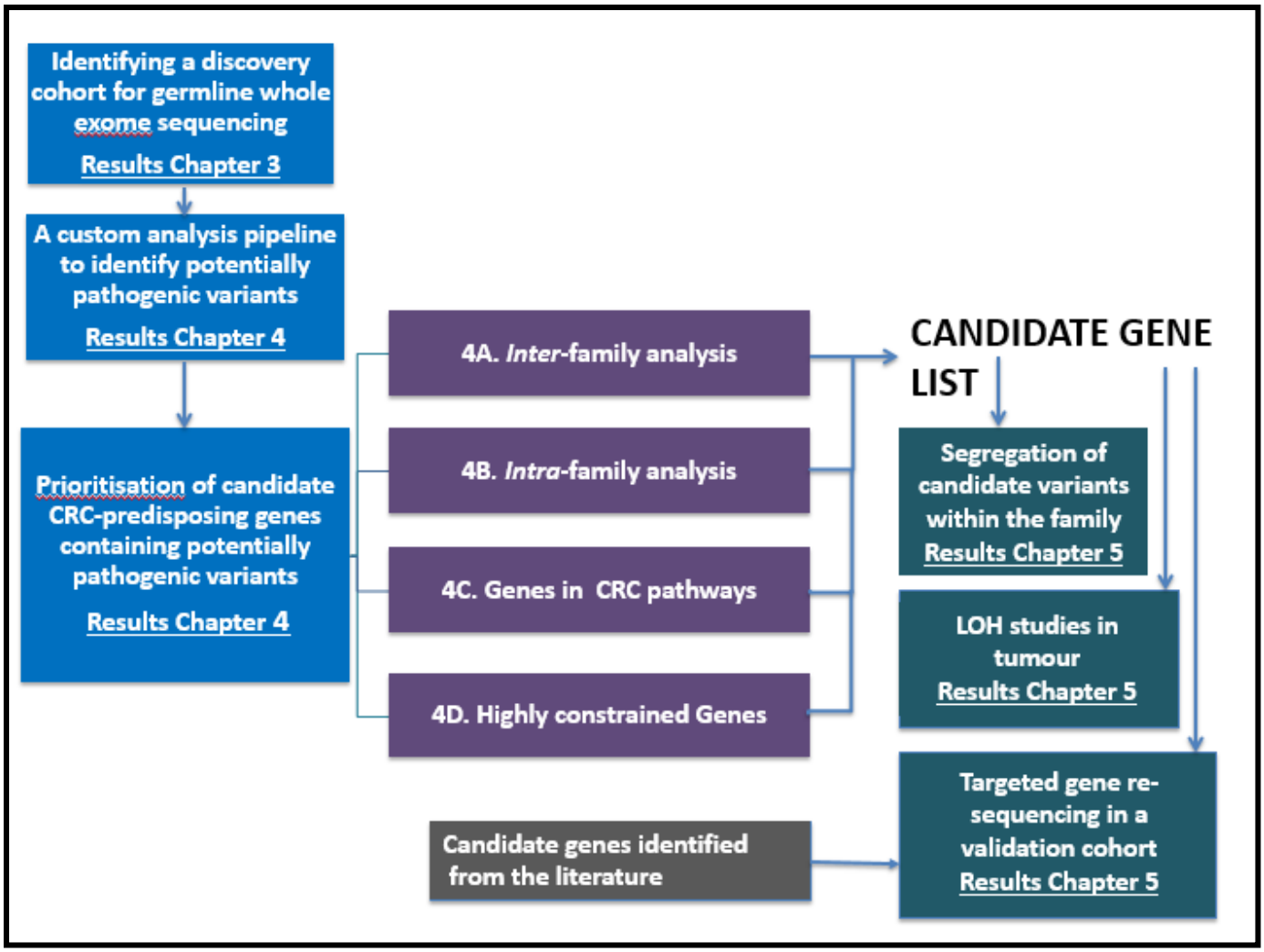



**Figure 1-1. Overview of thesis results chapters 3-5.**

CRC, colorectal cancer. LOH, loss-of- heterozygosity.

# Chapter 2. Materials and Methods

## 2.1. Materials

UltraPure DNase/RNase-Free Distillled Water was used in enzymatic reactions. Chemical solutions were made with water that was ultrafiltered using the Milli-Q Integral ultrafiltered water system.

### 2.1.1. Reagents and Chemicals

**Table 2-1. Reagents and Chemicals**

| Reagent | Supplier |
|---|---|
| 3130 POP-7 Polymer | Applied Biosystems (Foster City, CA, USA) |
| Agarose, LE, Analytical Grade | Promega (Madison, WI, USA) |
| Eosin 1% Aqueous | Australian Biostain (Traralgon, Vic, Australia) |
| Ethanol, Analytical Grade | Merck (Kilsyth, Vic, Australia) |
| Ethidium Bromide | Merck (Kenilworth, NJ, USA) |
| GelRed Nucleic Acid Stain, 10000X in water | Biotium (Fremont, CA, USA) |
| Haematoxylin | Australian Biostain (Traralgon, VIC, Australia) |
| Hi-Di Formamide | Thermo Fischer Scientific (Scoresby, VIC, Australia) |
| Loading Dye | Sigma-Aldrich (St. Louis, MO, USA) |
| $MgSO_4$ | BDH-Merck (Kilsyth, Vic, Australia) |
| Sodium Acetate | Merck (Kenilworth, NJ, USA) |
| UltraPure DNase/RNase-Free Distilled Water | Thermo Fischer Scientific (Scoresby, VIC, Australia) |
| Xylene | Biolab (Clayton, VIC, Australia) |

### 2.1.2. Enzymes and Buffers

**Table 2-2. Enzymes and Buffers**

| Enzyme or Buffer | Supplier |
|---|---|
| Exonuclease 1 | New England Biolabs (Ipswich, Massachusetts, USA) |
| HotStar Taq Polymerase | Qiagen (Valencia, CA, USA) |
| Proteinase K | Qiagen (Valencia, CA, USA) |
| RNAse A | Qiagen (Valencia, CA, USA) |
| Shrimp alkaline phosphatase | New England Biolabs (Ipswich, Massachusetts, USA |
| 10X sequencing buffer | Qiagen (Valencia, CA, USA) |

### 2.1.3. Nucleic Acids and Nucleotides

**Table 2-3. Nucleic Acids and Nucleotides**

| Nucleic Acids or Nucleotides | Supplier |
|---|---|
| dNTPs | Promega (Madison, WI, USA) |
| Human Genomic DNA:Female | Promega (Madison, WI, USA) |

### 2.1.4. Kits

**Table 2-4. Kits**

| Kit | Supplier |
|---|---|
| Agilent SureSelect Human All Exon v4 (51 Mb) | Agilent Technologies (Santa Clara, California, USA) |
| Agilent SureSelect Human All Exon (37.6 Mb) | Agilent Technologies (Santa Clara, California, USA) |
| | |
| Big Dye Terminator v3.1 Cycle Sequencing Kit | Applied Biosystems (Mulgrave, Victoria, Australia) |
| DNeasy Blood and Tissue Kit | Qiagen (Hilden, Germany) |
| Haloplex HS Custom Design Target Enrichment Kit 501 kb-2.5 Mb for Illumina | Agilent Technologies (Santa Clara, California, USA) |
| Qubit® dsDNA BR (Broad Range) Assay | Life Technologies (Carlsbad, CA, USA) |
| Qubit® dsDNA HS (High Sensitivity) Assay | Life Technologies (Carlsbad, CA, USA) |
| Repli-G Mini Kit | Qiagen (Hilden, Germany) |

### 2.1.5. Molecular Weight Standards

**Table 2-5. Molecular Weight Standards**

| Molecular Weight Standard | Supplier |
|---|---|
| 100 bp DNA ladder | Invitrogen (Carlsbad, CA, USA) |
| 1 Kb Plus DNA Ladder | Invitrogen (Carlsbad, CA, USA) |

### 2.1.6. Equipment

ABI 3130 Genetic Analyser (Applied Biosystems, Foster City, CA, USA)

Bravo NGS (Option A) (Agilent Technologies, Santa Clara, CA, USA)

Centrifuge 5415 R (Eppendorf, Hamburg Germany)

Corning LSE™ Vortex Mixer (Corning Inc, NY, USA)

Hybridisation Oven (Ratek, Boronia, VIC, Australia)

Megafuge 1.0R (Heraeus Instruments, Hanau, Germany)

Microcentrifuge SIGMA 1-15 (SIGMA Laboratory Centrifuges, Osterode am Harz, Germany)

Milli-Q Integral ultrafiltered water system (Merck Millipore, Bayswater, VIC, Australia)

Molecular Imager FX (Bio-Rad, Hercules, CA, USA)

Professional Thermocycler (Biometra, Göttingen, Germany)

Qubit® Fluorometer (Invitrogen, Carlsbad, CA, USA)

Stemi 2000 dissecting stereomicroscope, (Carl Zeiss, Oberkochen, Germany)

2200 TapeStation (Agilent Technologies, Santa Clara, CA, USA)

### 2.1.7. Software Packages

**Table 2-6. Software Packages**

| Software | Source |
|---|---|
| 3130 Data Collection version 3.0 | Applied Biosystems, Foster City, CA, USA |
| Geneious® v8.1.5 | Biomatters Ltd, Auckland, NZ |
| Progeny 9 | Progeny Genetics, LLC, Delray Beach, FL, USA |
| Microsoft Excel 2010 | Microsoft Corporation, Redmond, WA, USA |
| NEXUS Copy Number™ | Biodiscovery, Inc., El Segundo, CA, USA |
| Quantity One v4.6.9 | Biorad Laboratories, Hercules, CA, USA |

### 2.1.8. Custom Oligonucleotides

**Table 2-7. Primers for validation of germline variants**

| Gene | Sequencing Direction | Primer Sequence (5'-3') | Expected PCR Product Size (bp) |
|---|---|---|---|
| ***APLF*** | Forward | GAAGGGGCCTAATCCTTGC | 254 |
| ***APLF*** | Reverse | CGGCCCCAAACTTCTAAAAT | 254 |
| ***B4GALT1*** | Forward | GCGTCACCCTCGTTTACTACC | 249 |
| ***B4GALT1*** | Reverse | GGACCGAGGTCAAGTTGCTA | 249 |
| ***BCL9*** | Forward | CCCTGAGGAGATGCTGAAAT | 222 |
| ***BCL9*** | Reverse | ATATGACTGAGCCGGCTGTT | 222 |
| ***BOC*** | Forward | ACCAGCTACCTTCTCCAGCA | 243 |
| ***BOC*** | Reverse | GGCCAGTGGAACTTGCTTAG | 243 |
| ***C2CD3*** | Forward | ACAATTCCCAAAGCAGCAGT | 286 |
| ***C2CD3*** | Reverse | GGTCATTTCACCTTGGGAGA | 286 |
| ***CNTN6*** | Forward | CAACCTGGTCCAAGGAGAAA | 295 |
| ***CNTN6*** | Reverse | GGCCAGTTGAGGTTCTACCA | 295 |
| ***DMC1*** | Forward | CACAGGTACATTGTTTGGCTTC | 245 |
| ***DMC1*** | Reverse | TACATTGACCCAAGCATCCA | 245 |
| ***EP300*** | Forward | TGCATTCCCTGTGTCAAAAA | 284 |
| ***EP300*** | Reverse | TTGTGAGCATGCAAAAGGAG | 284 |
| ***ERCC3*** | Forward | TCCTACATGACGAGGGCTTT | 233 |
| ***ERCC3*** | Reverse | TATGCCACACAAACCTTGGA | 233 |
| ***GDF2*** | Forward | ACAGATGCCTGGGATAGTGC | 254 |
| ***GDF2*** | Reverse | CTGCTGTGGTCATTGGAGAA | 254 |
| ***INTU*** | Forward | CCAAGCGCTGCAATAAAAA | 218 |
| ***INTU*** | Reverse | TCTCCAGCTCCACTTGGTCT | 218 |
| ***JAG2*** | Forward | GTGTGGTGGACACGCAAG | 240 |
| ***JAG2*** | Reverse | ATCCTCGTCCTCCTCATCCT | 240 |
| ***LRRFIP2*** | Forward | TCGCTATGCAATCATTGAAGTT | 195 |
| ***LRRFIP2*** | Reverse | TGAGAGACAACAGAACTAACCAGT | 130 |
| ***MUS81*** | Forward | AGGTGTTCGAATCCCGACT | 290 |
| ***MUS81*** | Reverse | AGGGGTAGGGTACCTGGTCA | 290 |
| ***MYCBP2*** | Forward | GGGGGCCATATTTTTGTTTT | 239 |
| ***MYCBP2*** | Reverse | TTTTTCCATATGTGCCACCA | 239 |
| ***NOTCH4*** | Forward | ATGTCAGCAGTTTGCTGTGG | 291 |
| ***NOTCH4*** | Reverse | GTGTGTGCTAACCTGGCAGA | 291 |
| ***PHRF1*** | Forward | CACCTTCTTTGGCTCTGAGG | 178 |
| ***PHRF1*** | Reverse | CCTCTTCCTCTTGGCCTTCT | 178 |

PCR, polymerase chain reaction. bp, base pair.

**Table 2-7 (continued). Primers for validation of germline variants**

| Gene | Sequencing Direction | Primer Sequence (5'-3') | Expected PCR Product Size (bp) |
|---|---|---|---|
| ***POLN*** | Forward | TACAGACTGTGGTGAGCACT | 181 |
| ***POLN*** | Reverse | TCTTGTGAGCCTCTGTCAGC | 181 |
| ***SIN3B*** | Forward | CACCTCCCACATGTGTCCTTA | 230 |
| ***SIN3B*** | Reverse | CCTCTGTTCCCTGGTCACAT | 230 |
| ***SGPL1*** | Forward | AGTATGAGCCCTGGCAGCTA | 242 |
| ***SGPL1*** | Reverse | CACTCCAGCCTAGCAACAGA | 242 |
| ***TCF7L2*** | Forward | CTTCCTCCTGCCTTCTCCTT | 242 |
| ***TCF7L2*** | Reverse | CTGGCCTTGAATTTCTCCTG | 242 |
| ***USP28*** | Forward | GCTGCCAAATGCTGTTAAATC | 268 |
| ***USP28*** | Reverse | TACCCGAACCACAAATCTCC | 268 |
| ***WNT10A*** | Forward | ATGCCAACACAGTGTGCCTA | 264 |
| ***WNT10A*** | Reverse | CTGAGGTGGAGATGCTGGAT | 264 |
| ***XRCC6BP1*** | Forward | CCATGAACCTTTTGGAAGGA | 288 |
| ***XRCC6BP1*** | Reverse | TGCCTTGGAGTTTAAAGCAG | 288 |

PCR, polymerase chain reaction. bp, base pair.

**Table 2-8. Primers containing the M13F sequence for assessment of loss-of-heterozygosity in tumour tissue**

| *Gene* | Sequencing Direction | Primer Sequence | PCR Product Size (bp) |
|---|---|---|---|
| ***ASCL3*** | Forward | GTAAAACGACGGCCAGTTGTGTCAATGAAGGCTACGC | 91 |
| ***ASCL3*** | Reverse | CAGCTCTGAGGGTTTCCACT | 91 |
| ***DMC1*** | Forward | GTAAAACGACGGCCAGTTCCTTAATCTTGGCATACTCCAG | 86 |
| ***DMC1*** | Reverse | TCCAGTACTGCATCATGGTCT | 86 |
| ***ERCC3*** | Forward | GTAAAACGACGGCCAGTTTCTGTGGTCTAACACAGCTTCA | 107 |
| ***ERCC3*** | Reverse | TTTGGGGTTGTGCTTGAAAT | 107 |
| ***FKBP10*** | Forward | GTAAAACGACGGCCAGTTGGGGACTTTGTTCGATACC | 101 |
| ***FKBP10*** | Reverse | CACCCACCTGCCCAGTC | 101 |
| ***GALC*** | Forward | CTGGGGAAGAGCCATGGTAG | 95 |
| ***GALC*** | Reverse | GTAAAACGACGGCCAGTGTCACCTGGTGGAGCACTTT | 95 |
| ***GPR144*** | Forward | CTGCCAAGGAGCACAGC | 110 |
| ***GPR144*** | Reverse | GTAAAACGACGGCCAGTCTGGATCCTAGGGCATCAGG | 110 |
| ***IL4R*** | Forward | GTAAAACGACGGCCAGTCAGACCCCCTTGTGGACAG | 98 |
| ***IL4R*** | Reverse | ATCCTCCTGGCCATGACAC | 98 |
| ***JAG2*** | Forward | GAGAGCGCCAACAACCAGT | 106 |
| ***JAG2*** | Reverse | GTAAAACGACGGCCAGTGCGGCGTGAAGTTCTTG | 106 |
| ***LRRFIP2*** | Forward | GTAAAACGACGGCCAGTTCGCTATGCAATCATTGAAGTT | 130 |
| ***LRRFIP2*** | Reverse | TGAGAGACAACAGAACTAACCAGT | 130 |
| ***LRRK2*** | Forward | GTAAAACGACGGCCAGTGTGACAGCCAGATCATCAGC | 114 |
| ***LRRK2*** | Reverse | GACCAAGCCAAGAAGGTTCA | 114 |
| ***MMP2*** | Forward | AGCTCCCGGAAAAGATTGAT | 100 |
| ***MMP2*** | Reverse | GTAAAACGACGGCCAGTAGCTGAGGGCCAAGGAAG | 100 |
| ***MS4A15*** | Forward | GGTCATTGCCATGCACTTC | 105 |
| ***MS4A15*** | Reverse | GTAAAACGACGGCCAGTCACAGGCTCAGCTCAGGTTT | 105 |
| ***MUS81*** | Forward | GTAAAACGACGGCCAGTCTTCGTTCGCTGGCTGAC | 83 |
| ***MUS81*** | Reverse | GACCCACCTTCTGAAATACGA | 83 |
| ***PDE5A*** | Forward | GTAAAACGACGGCCAGTAAACGAACTTCCCACGTTTG | 86 |
| ***PDE5A*** | Reverse | AAGTGGGAAGGGACGTAGG | 86 |
| ***PHRF1*** | Forward | GTAAAACGACGGCCAGTCCTGTGTGACTGTCGTGGAG | 118 |
| ***PHRF1*** | Reverse | CGGGAGCTGGATGTGG | 118 |
| ***RIPPLY3*** | Forward | TACCCATGTCAAAGCGTCAA | 87 |
| ***RIPPLY3*** | Reverse | GTAAAACGACGGCCAGTAGAAGTCAATCGTGGCTTGC | 87 |
| ***SERPINB7*** | Forward | TTTTCCTCAGTTCAAGATAGAGAAGA | 112 |
| ***SERPINB7*** | Reverse | GTAAAACGACGGCCAGTCGAAGCAATCCCAGAGAGAT | 112 |
| ***SH2D4A*** | Forward | GTAAAACGACGGCCAGTGAAGCAGATTTGTAAGAGCTGGA | 83 |
| ***SH2D4A*** | Reverse | TGGAAGGTGCTATGGAGACC | 83 |
| ***SIN3B*** | Forward | GTAAAACGACGGCCAGTCCTCCCACATGTGTCCTTAGT | 105 |
| ***SIN3B*** | Reverse | GGGAGCGCTTCCTGCT | 105 |

PCR, polymerase chain reaction. bp, base pair.

**Table 2.8 (continued). Primers containing the M13F sequence for assessment of loss-of-heterozygosity in tumour tissue**

| *Gene* | Sequencing Direction | Primer Sequence | PCR Product Size (bp) |
|---|---|---|---|
| ***SUCLG2*** | Forward | GGTGGTGTAAAGGAAGCTCAA | 114 |
| ***SUCLG2*** | Reverse | GTAAAACGACGGCCAGTTGACGGTGAGAAATTAAAACTGG | 114 |
| ***TEX14*** | Forward | GTAAAACGACGGCCAGTGCATCACTATCGGCACATTG | 90 |
| ***TEX14*** | Reverse | AAGAAACTGGGAGGTCAGCA | 90 |
| ***TMCC3*** | Forward | GTAAAACGACGGCCAGTCGGGACGTATCAAGCAAGTC | 112 |
| ***TMCC3*** | Reverse | TGCTCGATCTCTCTGAGCTTT | 112 |
| ***UFL1*** | Forward | GTAAAACGACGGCCAGTCAGCTTGTGTTGGTCAAGGA | 85 |
| ***UFL1*** | Reverse | GCAATATCAACCCATGTTCCA | 85 |
| ***ZNF788*** | Forward | GAGCGGCCTTCAGGTACA | 119 |
| ***ZNF788*** | Reverse | GTAAAACGACGGCCAGTTGGTGTGGATCCTTTTATGATTT | 119 |
| ***ZNF788*** | Forward | GTAAAACGACGGCCAGTAAGATACGTGGAAAAGTCTTTCAT | 128 |
| ***ZNF788*** | Reverse | CGTTGAAAGGCATTGAAATAAA | 128 |

PCR, polymerase chain reaction. bp, base pair.

**Table 2-9. M13F primer for Sanger Sequencing Reaction using DNA from FFPE Tumour Tissue**

| Primer Name | Primer Sequence |
|---|---|
| **M13F** | GTAAAACGACGGCCAGT |

### 2.1.9. Tissue Specimens and Clinical Information

Each study participant gave informed consent for the collection of tissue specimens and clinical information. Whole blood or its derivatives, or saliva was collected in order to extract germline DNA. Colorectal cancer (CRC) tumour blocks or cut sections mounted on glass slides were requested from pathology laboratories for the purpose of loss-of-heterozygosity studies. Clinical information consisting of three generation pedigrees, histopathology reports, diagnostic sequencing reports for known CRC-predisposing genes and clinical letters were obtained from Familial Cancer Centre (FCC) files and medical records.

#### *2.1.9.1. Study Ethics Approval and Patient Recruitment Sites*

This study was approved by the Human Research and Ethics Committees (HRECs) of the following institutions: 1) Royal Melbourne Hospital (VIC, Australia) (multi-site approval that included approval for Monash Medical Centre (VIC, Australia) and Prince of Wales Hospital (NSW, Australia)); 2) Peter MacCallum Cancer Centre (VIC, Australia) and 3) King Edward Memorial Hospital (WA, Australia). Management of ethics amendments and drafting of the HREC application for the King Edward Memorial Hospital site was performed by myself.

**Table 2-10. HREC Sites and Study Numbers**

| HREC Site | HREC Study Number | Principal Investigators |
|---|---|---|
| **Royal Melbourne Hospital** | 12/MH/145 | Associate Professor Alison Trainer, Associate Professor Lara Lipton |
| **Peter MacCallum Cancer Centre** | 10_83 | Professor Ian Campbell, Associate Professor Alison Trainer, Associate Professor Paul James |
| **King Edward Memorial Hospital** | 2013087EW | Professor Ian Campbell, Associate Professor Alison Trainer, Associate Professor Paul James |

Participants were enrolled into the study from FCCs within each institution. Individuals enrolled onto the study through the FCCs who were not included in the discovery cohort were entered into the validation cohort. For the validation cohort, the cut-off age for isolated CRC cases was increased to a diagnosis age of 50 years or younger.

Germline DNA was also obtained from patients in the Molecular and Cellular Oncology (MCO) Group Study who had given prior consent for genetic research studies through Professor Robyn Ward from the University of New South Wales Lowy Cancer Research Centre (NSW, Australia) (Ward and Hawkins, 2013, Sloane et al., 2014). The MCO study prospectively enrolled more than 1500 patients who underwent surgical resection for CRC between 1994 and 2010. Germline DNA from patients who had developed CRC prior to the age of 50 years, and who had consented for genetic research studies, was obtained for this study. A family history of CRC was not required for the MCO patients due to the young onset of CRC. Self-reported family history was obtained from the MCO database. Results of testing for MMR

proteins by IHC was extracted from the MCO database and provided in table form. Patients with a known pathogenic variant in a CRC-predisposing gene were excluded.

Additional germline DNA for the validation cohort was obtained from the Victorian Cancer Biobank (VCB) (VIC, Australia) and the Kathleen Cuningham Foundation Consortium (KConFab) (www.kconfab.org). Applications to access biospecimen germline DNA were submitted to VCB and kConFab. Individuals who consented to their germline DNA being banked by VCB and kConFab had already consented to genetic research being performed on their tissue at the time of tissue collection.

Germline DNA was requested for individuals who developed CRC aged $\leq$ 50 years, regardless of family history. Where available, pathology reports for the individual's CRC were reviewed for MMR IHC results, and all cases with MMR deficiency on IHC were excluded.

VCB provided a buffy coat for each individual. DNA was extracted as described in Chapter 2 Section 2.2.1. kConFab provided 500 ng of germline DNA per individual, obtained from EBV-transformed lymphoblastoid cell lines.

#### *2.1.9.2. Germline DNA*

Twenty millilitres (mLs) of whole blood were collected in ethylenediaminetetraacetic acid tubes from patients at local pathology centres prior to being sent to the Molecular Pathology department at Peter MacCallum Cancer Centre (PMCC) where DNA extraction was performed from peripheral blood leucocytes. For patients enrolled outside Victoria, germline DNA from peripheral blood leucocytes or saliva was extracted prior to being sent to the Cancer Genetics Laboratory at PMCC. In instances where germline DNA was available from previous diagnostic sequencing studies performed through the FCCs, this was retrieved for the study. Two to three µg of DNA were requested for each patient sample. VCB samples were provided as cell pellets or buffy coats and germline DNA extracted by the Molecular Pathology Department at PMCC, apart from 71 samples that were extracted manually in the Campbell laboratory using the method described in 2.2.1.

#### *2.1.9.3. Tumour blocks or sections*

Formalin-fixed, paraffin-embedded (FFPE) tumour blocks were requested from the relevant pathology laboratories. The pathology department at PMCC cut 10 x 5 µm sections from each tumour block, and mounted them on glass slides.

## 2.2. Methods

### 2.2.1. DNA Extraction from Buffy Coats or Cell Pellets Derived from Human Blood

DNA extraction was performed using the DNeasy Blood and Tissue Kit according to the manufacturer's instructions, apart from the final DNA elution step that was performed using a reduced volume (50 µL instead of 200 µL) of Buffer AE to increase the final DNA concentration. Four µL of RNAse A (100 milligram/mL) were added at Step 1 as per protocol.

### 2.2.2. DNA Quantification

DNA was quantified by fluorescent dye tagging, using Qubit® dsDNA HS Assay kits (for estimated DNA concentrations of <100 ng) or Qubit® dsDNA BR Assay kits (for estimated DNA concentrations of ≥100 ng), and a Qubit® Fluorometer as per the manufacturer's instructions.

### 2.2.3. Gel Electrophoresis

Agarose gel electrophoresis was used to determine the size of 1) DNA fragments in extracted germline DNA prior to DNA for WES (2.2.4) or targeted gene sequencing (2.2.13), and 2) polymerase chain reaction (PCR) products for Sanger sequencing (2.2.9.1). Gels containing 0.8% agarose (for genomic DNA assessment) or 2% agarose (for PCR product assessment) in 1 X tris-acetate- ethylenediaminetetraacetic acid (TAE) buffer were made, with 5 µl of ethidium bromide or GelRed Nucleic Acid Stain added to visualise DNA under ultra-violet light. Two µL of loading dye were added to each DNA aliquot prior to being loaded into gel wells. Three µL of the appropriate range DNA size ladder were loaded in a separate gel well. Gels were placed in 1 x tris-acetate- ethylenediaminetetraacetic acid and electrophoresed at 100-120 volts for 15-30 minutes. DNA within the electrophoresed gels was visualised on a Molecular Imager FX using Quantity One software at a scanning resolution of 100 µm. Genomic DNA samples that did not have a high molecular weight band above the 12 kb marker were omitted from whole exome or targeted sequencing.

### 2.2.4. WES of Germline DNA

One µg of germline DNA extracted from blood was submitted to BGI Tech Solutions (Hong Kong, China) or the Molecular Core Facility of PMCC for WES using either the Agilent SureSelect Human All Exon v4 (51 Mb) or Agilent SureSelect Human All Exon (37.6 Mb) for exome enrichment. Paired-end 100 bp reads were sequenced on an Illumina Hiseq 2000. The mean coverage of target bases across the 54 exomes was 83X (range 49-136X).

### 2.2.5. Sequence Alignment and Variant Calling from WES Reads

Alignment of sequenced reads and variant calling was performed by the Research Computing Facility, PMCC using bioinformatics pipelines assembled from publicly available software, together with custom scripts.

FASTQC (http://www.bioinformatics.babraham.ac.uk/projects/fastqc/) was used to check raw sequence data quality. FASTQC runs 11 modules on FASTQ files to assess quality (basic statistics, per base sequence quality, per sequence quality scores, per base sequence content, per base GC content, per sequence GC content, per base N content, sequence length distribution, sequence duplication levels, overrepresented sequences and kmer content), with the majority of outputs presented in graph form. FASTQC outputs were analysed to ensure that the Illumina sequencing data was consistent with exome sequencing performed on human samples. Low quality bases (quality score <20) and adaptors were trimmed using CutAdept (Martin, 2011). Paired-end 100 bp reads were aligned to the reference human genome (NCBI build 37, hg19) using BWA-MEM. Duplicate reads were marked using Picard (http://broadinstitute.github.io/picard). BAM files of multiple individuals from the same family were merged for joint genotyping.

Local realignment around indels and base quality score recalibration was performed using Genome Analysis Toolkit (GATK) v3.1 (McKenna et al., 2010). GATK HaplotypeCaller was used to identify single nucleotide variants (SNVs) and indels. Variant annotation from Ensembl release 73 was performed using Ensembl's Perl API and Variant Effect Predictor (McLaren et al., 2010). Variants were also annotated with Combined Annotation-Dependent Depletion (CADD) scores (Kircher et al., 2014) and variant frequencies from the Non-The Cancer Genome Atlas (TCGA) VCF Subset (ExAC Browser beta 2014) of the Exome Aggregation Consortium (ExAC) dataset (Lek et al., 2016b), using custom Python scripts.

### 2.2.6. Software used for analysis of germline variants in WES and targeted gene sequencing data

Variant filtering was performed using the Galaxy platform (Giardine et al., 2005) and Microsoft Excel 2010, while variant analysis was performed using Microsoft Excel 2010. Variant filtering and analysis is explained in detail in Chapters 4 to 5.

### 2.2.7. Primer Design

Primers were designed using an online primer design site (http://bioinfo.ut.ee/primer3-0.4.0/primer3/). The product size range was 100-300 bases. Default parameters were used, including a primer size of 18-27 bases, minimum, optimal and maximum melting temperatures of 57°C, 60°C and 63°C, and a primer GC percentage of between 20-80%. Primer sequences were checked for binding specificity using the University of California, Santa Cruz In-Silico PCR tool (http://www.genome.ucsc.edu/cgi-bin/hgPcr). Primers that were predicted to bind exclusively to the targeted region of the human genome were selected.

For primers used for PCR amplification of tumour DNA, M13 forward (M13F) tailed primers were designed. If the forward strand had a mononucleotide repeat of 5 or more bases 5' to the region of interest, the M13F primer was placed on the reverse primer, and the reverse sequence was checked to ensure there were no mononucleotide repeats of 5 or more bases preceding the region of interest, in order to minimise sequencing errors due to polymerase slippage.

### 2.2.8. Whole Genome Amplification

Twenty to twenty-five ng of germline DNA were used for whole genome amplification using the REPLI-g Mini Kit as per manufacturer's instructions. Whole genome amplification was performed on germline DNA due to the limited amount of germline DNA obtained for the study. Only unamplified germline DNA was sent for exome sequencing, while validation of variants identified in sequenced exomes was performed on whole genome amplified DNA using Sanger sequencing.

### 2.2.9. Sanger Sequencing for Validation of Variants Detected in Next Generation Sequencing data

#### *2.2.9.1. Optimisation of PCR Conditions for Primer Pairs*

Optimisation of PCR conditions for each primer pair was carried out. For each primer pair, a 10 µL PCR reaction was made containing 1 µL of PCR buffer, 0.5 µL of 4 µM dNTPs, 0.05 µL of HotStar Taq DNA Polymerase, 1 µL of the primer pair at a 1 µM concentration, 10 ng of Human Genomic DNA: Female DNA and UltraPure DNase/RNase-Free Distilled Water. PCR mastermixes containing all components of the PCR reaction except for the primer pair and DNA template were prepared, and added to each 0.2 mL PCR tube containing the primer pair and DNA template.

A touchdown PCR program was used to increase PCR product specificity. PCR conditions consisted of an activation step at 95°C for 15 minutes, 10 amplification cycles consisting of 95°C for 30 seconds, 65°C (decreasing by 1°C every cycle) for 30 seconds, and 72°C for 30 seconds, 35 amplification cycles consisting of 95°C for 30 seconds, 55°C for 30 seconds and 72°C for 30 seconds, and an elongation step at 72°C for 10 minutes.

Four µL of PCR product were separated on a 2% agarose gel using agarose gel electrophoresis, as described above. Three µL of 100 bp DNA ladder were used to determine the size of the amplified PCR products. If no PCR product was seen, the initial touchdown PCR was repeated with the addition of 1 µL of 2 millimole $MgCl_2$. If multiple bands were seen on the agarose gel, further optimisation was performed using a temperature gradient PCR program, using a second PCR reaction made as described above, with the following PCR conditions: an activation step at 94°C for 15 minutes, 40 amplification cycles consisting of 94°C for 30 seconds, a temperature gradient of 55.0°C, 55.4°C, 56.5°C, 58.0°C, 59.8°C, 61.6°C, 63.4°C, 65.2°C, 67.0°C, 68.5°C, 69.6°C and 70.0°C for 30 seconds, and 72°C for 30 seconds, and an extension step at 72°C for 10 minutes. PCR products were visualised on a 2% agarose gel as

described above, and PCR reaction conditions that yielded a single band at the expected size were used for amplification of sample DNA template. Once PCR conditions had been optimised for each primer pair, whole genome amplified DNA from the appropriate patient sample was used as DNA template for PCR amplification and subsequent Sanger sequencing.

#### *2.2.9.2. PCR Product Clean-Up*

Left-over primers and dNTPs in the PCR reactions were removed using Exonuclease 1 and shrimp alkaline phosphatase. A mastermix of 1 µL of shrimp alkaline phosphatase, 0.25 µL of Exonuclease 1 and 1.75 µL of UltraPure $H_20$ per PCR reaction was made and kept on ice. Three µL of mastermix were added to 6 µL of PCR reaction prior to placement in a thermocycler at 37°C for 15 minutes, followed by 80°C for 15 minutes to inactivate the enzymes.

#### *2.2.9.3. Big Dye Terminator (BDT) v3.1 Cycling Sequencing*

Four µL of each PCR product post clean-up were added to 1 µL of BDT v3.1 Sequencing Mix, 4 µL of a 100 nanomole forward or reverse primer, 3.5 µL of 5X Big Dye Sequencing Buffer and 7.5 µL of UltraPure DNase/RNase-Free Distilled Water to make up a 20 µL BDT sequencing reaction. Sequencing was performed using 25 cycles of the following conditions: 96°C for 10 seconds, 50°C for 5 seconds, 60°C for 4 minutes.

#### *2.2.9.4. DNA Precipitation With Ethanol*

DNA precipitation was performed in 1.5 mL micro-centrifuge tubes (for up to 24 samples) or a 96 reaction well plate (for more than 24 samples) to remove unincorporated dye terminators that could interfere with sequence detection during capillary electrophoresis.

##### 2.2.9.4.1. DNA Precipitation using 1.5ml microcentrifuge tubes

Sixteen µL of UltraPure DNase/RNase-Free Distilled Water were first added to each sequencing reaction. Each sequencing reaction was then mixed with 10 µL of 3 M sodium acetate (pH 5.2) and 54 µl of 100% analytical grade ethanol (EtOH) in 1.5 mL microcentrifuge tubes. Reactions were placed on ice for 15 minutes. Sequencing reactions were then centrifuged at 13000 revolutions per minute (rpm) at 4°C for 20 minutes in a tube centrifuge. The supernatant was carefully removed.

Two hundred and fifty µL of 70% EtOH were added to the DNA pellet to remove excess sodium acetate. Each sample was vortexed using a vortex mixer prior to being centrifuged at 13000 rpm at room temperature for 5 minutes. The supernatant was removed and DNA pellets were dried in a hybridisation oven at 56°C for 5 minutes to remove residual EtOH.

##### 2.2.9.4.2. DNA Precipitation using a 96 reaction well plate

Post BDT sequencing, the reactions were transferred to a 96 reaction well plate. One point five µL of 3M sodium acetate (pH 5.2) and 38.5 µL of 100% analytical grade EtOH were added to each reaction. The plate was sealed and inverted to mix the contents of each well, followed by incubation on ice for 15 minutes.

Each plate was centrifuged at 2800 rpm for 40 minutes at 4°C in a plate centrifuge. The plate was then unsealed and inverted over a paper towel to remove the supernatant. The inverted plate was then spun at 600 rpm for 1 minute at 4°C. One hundred and fifty µL of the 70% EtOH which had been placed in ice were added to each sample. The plate was then sealed, vortexed

and centrifuged at 2800 rpm for 5 minutes at 4°C. The plate was then unsealed and inverted over a paper towel. The inverted plate was centrifuged at 300 rpm for 2 minutes at 4°C. The plate was then uncovered and placed upright in a hybridisation oven at 50°C for 5 minutes to remove residual EtOH.

#### *2.2.9.5. Capillary Electrophoresis*

Fifteen µL of Hi-Di formamide were added to resuspend the DNA pellet. If microcentrifuge tubes were used for DNA precipitation, the resuspended samples were transferred to a 96 reaction well plate for capillary electrophoresis. Capillary electrophoresis was performed on the ABI 3130 Genetic Analyser according to a standard sequencing method instrument protocol operated by 3130 Data Collection version 3.0 software, using POP-7 Polymer and a 36 centimetre capillary array. Sequence traces were analysed using Geneious® software version 8.1.5.

### 2.2.10. DNA Extraction From FFPE Tumour Tissue

Unstained tumour sections were stained with haematoxylin and eosin (H&E) prior to micro-dissection of tumour tissue and DNA extraction. Staining with H&E (Bancroft 2013) was performed as described below, using a slide rack and glass or plastic staining containers containing sufficient solution to completely immerse all tissue sections.

#### *2.2.10.1. Preparation of Scott's Tap Water for Use During H&E Staining*

A magnetic stirrer was placed in a 2 L plastic beaker and the beaker was filled with 1.2 L of tap water. Forty milligrams of anhydrous magnesium sulphate were added to the beaker. Four g of potassium bicarbonate were then added to the beaker. The mixture was stirred till both chemicals had completely dissolved. Tap water was then added to a final volume of 2 L. The solution was briefly stirred prior to the addition of 0.05 g of thymol. The solution was then stirred for 1 hour.

#### *2.2.10.2. H&E Staining of Tumour Sections for Micro-dissection*

The FFPE tumour sections were deparaffinised, hydrated, stained with H&E, and dehydrated as shown in the 14 steps below:

**Table 2-11. H&E staining protocol**

| Step | Stain/Wash | Duration |
|---|---|---|
| 1 | Xylene | 4 minutes |
| 2 | Xylene | 4 minutes |
| 3 | 100% EtOH | 1 minute |
| 4 | 70% EtOH | 1 minute |
| 5 | Water | 1 minute |
| 6 | Haematoxylin | 2 minutes |
| 7 | Water | 30 seconds |
| 8 | Water | 30 seconds |
| 9 | Scott's Tap Water | 30 seconds |
| 10 | Water | 30 seconds |
| 11 | 1% Eosin | 15 seconds |
| 12 | 100% EtOH | 4 dips |
| 13 | 100% EtOH | 1 minute |
| 14 | 100% EtOH | 1 minute |

Stained slides were air-dried for 10 minutes prior to micro-dissection.

#### *2.2.10.3. Tumour Micro-dissection and DNA Extraction*

The first H&E slide of each tumour block was reviewed by a gastro-intestinal pathologist at the Department of Pathology, PMCC who confirmed the presence and percentage of tumour within each section. Up to 9 H&E stained sections of each tumour were micro-dissected by hand using sterile 23 gauge needles. Each micro-dissected tumour was placed in a 1.5 mL microcentrifuge tube containing 100 µL of Buffer ATL from the DNeasy Blood and Tissue Kit. DNA extraction was performed using the DNAeasy Blood and Tissue Kit according to the manufacturer's instructions, apart from the final DNA elution step that was performed using a reduced volume (20 µL instead of 200 µL) of Buffer AE. Proteinase K digestions were incubated at 56°C in a hybridisation oven.

### 2.2.11. Copy Number Analysis of Tumour DNA

Extracted tumour DNA was quantified using the Qubit fluorescence method (2.2.2). Thirty-seven to seventy ng of tumour DNA were submitted to Affymetrix (Thermo Fischer Scientific, Santa Clara, CA, USA) for genome-wide copy number analysis using the Oncoscan® FFPE assay kit, which utilises Molecular Inversion Probe technology (Wang et al., 2012) optimised for degraded FFPE samples. Copy number data were analysed using NEXUS Copy Number™ software.

### 2.2.12. Ascertainment of germline variants that fall in regions of loss of heterozygosity (LOH) in tumour DNA

A custom Python script was used to identify filtered germline variants (Chapter 4) that were located in regions of LOH in corresponding tumour DNA. The output from this script was analysed in Microsoft Excel 2010.

### 2.2.13. Targeted Gene Sequencing

Sequencing libraries were prepared from genomic DNA using the Haloplex HS Custom Design Target Enrichment Kit 501 kb-2.5 Mb for Illumina (Agilent Technologies), according to the Haloplex$^{HS}$ Target Enrichment System Automated Protocol for Illumina sequencing. Sequencing was performed by the Australian Genome Research Facility (North Melbourne, VIC, Australia) on an Illumina Hiseq 2500. Ninety-six samples were run in each lane, with an expected mean coverage of 441X.

#### *2.2.13.1. Designing the Customised Gene Panel*

A customised Haloplex$^{HS}$ targeted gene panel was designed using Agilent's SureDesign tool (https://earray.chem.agilent.com/suredesign/index.htm). Probes were selected to tile the coding regions of candidate CRC-predisposing genes plus 10 flanking base pairs on either side of each exon. For established CRC-predisposing genes, the untranslated regions were included. Probes were first selected using the 'Maximise Specificity' parameter, as this minimises the capture of off-target regions. If less than 90% of a gene was covered using this parameter, the 'Balanced' parameter was used instead to increase gene coverage to more than 90%. Probes were grouped according to their parameters, and the final design was made by combining all the probe groups.

#### *2.2.13.2. Preparation of DNA Samples*

Genomic DNA samples were quantified using Qubit® dsDNA BR Assay as described above. DNA samples with a concentration of more than 60 ng/µL were diluted 2-5 fold using UltraPure DNase/RNase-Free Distilled Water to a concentration of less than 60 ng/µL. Diluted samples were re-quantified using Qubit® dsDNA BR Assay. DNA size was determined by agarose gel electrophoresis.

Each DNA sample was placed in a single well (excluding well A1) of a full-skirted 96-well Eppendorf twin.tec plate at a final DNA concentration of 2 ng/µL in a total volume of 35 µL of 10 millimole Tris buffer (pH 8.5). Thirty-two µL of Enrichment Control sample provided in the Haloplex$^{HS}$ kit were placed in position A1 of each full-skirted 96-well Eppendorf twin.tec plate.

#### *2.2.13.3. DNA Digestion*

The Restriction Enzyme (RE) Master Mix Source Plate was prepared using 16 different restriction enzymes, RE Buffer and BSA Solution provided in theHaloplex$^{HS}$ kit, as per protocol. Briefly, the 16 restriction enzymes were divided into 8 unique mastermixes, each containing 2 of the 16 restriction enzymes, RE Buffer and BSA Solution. Using the NGS Bravo Option A and VWorks software, each sample of DNA was divided among 8 wells in a 384 well reaction plate, and combined with the RE Master Mix. The restriction enzyme double-digests were carried out at 37°C for 30 minutes in a thermal cycler. The Enrichment Control sample was run on a 2200 TapeStation using the High Sensitivity D1000 ScreenTape to ensure that digestion had occurred.

#### *2.2.13.4. Probe Hybridisation and Sample Indexing*

A Hybridisation Master Mix containing custom HaloPlex$^{HS}$ biotinylated probes was prepared as per protocol. Indexing Primers were prepared in a half-skirted 96 –well Eppendorf twin.tec

plate. The eight digestion reactions for each genomic DNA sample, the hybridisation Master Mix and the Indexing Primers were combined using the NGS Bravo. Hybridisation was performed in a thermal cycler using a heated lid, at 95°C for 5 minutes, followed by 58°C for 16 hours overnight.

#### *2.2.13.5. Purification and Ligation of Probe to Circularised DNA Fragments*

Forty-five mL of freshly prepared 70% EtOH was placed in a Thermal Reservoir for washing of DNA. A Nunc Deepwell Source plate containing hybridisation stop solution and AMPure XP beads was prepared for removal of hybridisation solution, as per protocol.

A DNA ligation mastermix was prepared with HS Ligation Solution and 1 millimole rATP as per protocol. The 96-well Eppendorf twin.tec plate containing the hybridised samples was transferred to the NGS Bravo, that was then used for the liquid handling steps in DNA purification and addition of ligation mastermix. Ligation was performed according to manufacturer's instructions at 55°C for 10 minutes in a thermal cycler using a heated lid.

#### *2.2.13.6. Capture of ligated DNA/probe hybrids using streptavidin beads*

The 96-well Eppendorf plate containing the ligated DNA products was transferred to the NGS Bravo as per protocol. The circularised DNA bound to the biotinylated HaloPlex HS probes were captured using Streptavidin beads. The captured DNA-probe hybrids were washed with 1M sodium hydroxide. The liquid handling steps were performed by the NGS Bravo.

#### *2.2.13.7. PCR Amplification of DNA Libraries*

A PCR Mastermix source plate was prepared as per protocol. The NGS Bravo was used to add the PCR Mastermix to the captured DNA-probe hybrids. The prepared PCR plate was then sealed and placed in a thermal cycler. PCR amplification was performed using the following conditions: an activating step at 98°C for 2 minutes, 30 amplification cycles at 98°C for 30 seconds, 60°C for 30 seconds, and 72°C for 1 minute, and an extension step at 72°C for 10 minutes. The PCR amplified products were held at 8°C.

#### *2.2.13.8. DNA Purification of Amplified Libraries*

The amplified DNA was purified by the addition of AMPure XP beads , washed with 70% EtOH, and eluted using HS Elution Buffer, following the protocol. The NGS Bravo was used for the liquid handling steps.

#### *2.2.13.9. Quantification of Enriched Target DNA*

Sample quantification was performed on the 2200 TapeStation, using the High Sensitivity D1000 ScreenTape and reagent kit, as per the instrument user manual. The electropherogram for each sample was checked for a peak fragment size between 225 and 540 base pairs, and the concentration of enriched DNA was determined by integration under the peak between 175 and 625 base pairs.

#### *2.2.13.10. Sample Pooling for Multiplexed Sequencing*

Using the sample concentrations obtained from the 2200 TapeStation, a 500 µL pooled solution containing 40 nanomole concentrations of each of the 96 enriched target DNA samples was made, with the addition of UltraPure DNase/RNase-Free Distilled Water. Each pooled solution of 96 DNA samples was submitted to Australian Genome Research Facility (North Melbourne, VIC, Australia) for sequencing on a lane of the HiSeq 2500.

### *2.2.13.11. Alignment and Variant Calling from Haloplex HS Targeted Sequencing Reads*

Alignment of sequenced reads and variant calling was performed at the Research Computing Facility, PMCC using bioinformatics pipelines assembled from publicly available software, together with custom scripts.

Removal of sequenced read duplicates was performed using a custom script. FASTQC was used to check raw sequence data quality. Low quality bases and adaptors were trimmed using CutAdept. Sequences were aligned to the reference human genome (NCBI build 37, hg19) using BWA-MEM. Local realignment around indels and base quality score recalibration was performed using GATK v3.1. Alignment performance summaries were generated using SAMtools (Li et al., 2009) and custom scripts. Identification of SNVs and indels was performed using three variant callers, GATK Unified Genotyper, GATK Haplotype Caller and Platypus. Variant Call Format (VCF) files from all three callers were combined into a single VCF file using GATK. Variant annotation from Ensembl release 73 was performed using Ensembl's Perl API and Variant Effect Predictor. Variants were also annotated with CADD scores and variant frequencies from the Non-TCGA VCF Subset (ExAC Browser Beta 2014) of the Exome Aggregation Consortium (EXAC) dataset using custom Python scripts.

# Chapter 3. Preparation of a Discovery Cohort for the Identification of Novel CRC-Predisposing Genes

## 3.1. Introduction

This chapter presents the process of identifying individuals with an increased likelihood of harbouring an inherited rare, pathogenic variant predisposing to CRC or polyposis. These individuals will form a discovery cohort for identifying novel CRC-predisposing genes through WES of germline DNA. In Chapter 1, the clinical features of the known familial CRC syndromes were described. Individuals with CRC due to a germline predisposition often have distinct clinical characteristics compared to sporadic CRC. Careful selection of individuals is required in order to increase the likelihood of identifying novel CRC-predisposing genes, as including patients with a low likelihood of a high-risk inherited pathogenic variant would increase the noise-to-signal ratio in subsequent analyses.

All individuals were recruited through the familial cancer centres due to their increased familial risk. No pathogenic variants in known CRC-predisposing genes had been identified in these patients by clinical testing through the FCC. Inclusion criteria, and the resulting clinical and histopathologic characteristics of the discovery cohort are described in this chapter. The later analysis for potential pathogenic germline variants in novel genes is described in Chapter 4.

### 3.1.1. Clinical characteristics of individuals with an increased likelihood of a germline predisposition to CRC

#### *3.1.1.1. Family History of CRC*

A family history of CRC can be an indicator of a Mendelian genetic cause for the predisposition to CRC. As described in Chapter 1, the Amsterdam Criteria were initially developed as a research tool to identify high risk families with a likely monogenic familial predisposition to CRC. While many individuals with a family history of CRC do not meet the Amsterdam criteria, the risk of developing CRC on a background of a family history of CRC is increased, as described below.

The risk of developing CRC for an individual with a first degree relative (FDR) with CRC was evaluated in a meta-analysis of 26 case-control and cohort studies (Johns and Houlston, 2001). This meta-analysis found the relative risk to be 2.25 (95% CI 2.00-2.53), increasing to 4.25 (95% CI 3.01-6.08) for individuals with more than one relative with CRC. For individuals with a relative diagnosed with CRC before the age of 45 years, the estimated relative risk was 3.87 (95% CI 2.40-6.22). While these increased risks may be due to shared environmental as well as shared genetic factors, particularly if both affected individuals are from the same generation, a group of individuals with a family history of CRC is likely to be enriched for individuals with a germline predisposition to CRC. This likelihood increases if there are multiple individuals within a family affected with CRC or if a relative develops CRC before age 45 years. Families with multiple generations affected by CRC have an increased likelihood of a monogenic, dominant predisposition to CRC. A caveat on family history is that it may not be evident for some individuals, due to small family size, estrangement from biological family, or *de novo* germline

events (for example, 25% of germline *APC* mutations are estimated to be *de novo* (Bisgaard et al., 1994)).

#### *3.1.1.2. Young age of onset of CRC*

The average age at diagnosis of CRC in Australia is 69 years (Young et al., 2015) with only 6.2% occurring before the age of 50 years (Boyce et al., 2016) and 1.6% before age 40. Individuals with a germline predisposition to CRC often develop CRC at a younger age compared to the general population (Lynch and de la Chapelle, 2003). This is explained by Knudsen's two hit hypothesis that postulated that mutations in both alleles of a tumour suppressor gene in the same cell are required for the development of cancer. As the required 'first hit' to a tumour suppressor gene is already present in the germline of individuals with an inherited predisposition to CRC, these individuals only need one additional somatic mutation in the same gene to occur by chance, compared to individuals with sporadic CRC who require 2 somatic mutations to occur in that tumour suppressor gene. This is demonstrated by the young age of onset of CRC in patients with hereditary CRC such as FAP and Lynch Syndrome.

#### *3.1.1.3. Colonic Polyposis*

3.1.1.4. Colonic polyposis is a well-recognised phenotype associated with inherited pathogenic variants in CRC-predisposing genes, including *APC, MUTYH, POLE, POLD1, STK11, SMAD4, and BMPR1A.* The polyp types in colonic polyposis that are associated with CRC include adenomatous polyps, hamartomatous polyps, serrated polyps and mixed polyps.

##### *3.1.1.4.1. Adenomatous Polyposis*

Colonic adenomas are recognised precursors of CRC, and remnants of colonic adenoma are sometimes found within CRCs. The adenoma-to-carcinoma model proposed by Vogelstein (Vogelstein et al., 1988) described genetic alterations in tumour suppressor genes and oncogenes that accumulate as adenomas progress to CRCs. As colonic adenomas are a recognised precursor of CRC and a distinctive feature of hereditary CRC, individuals with 20 or more adenomatous polyps in the colon were included in the discovery cohort.

##### *3.1.1.4.2. Serrated and mixed polyposis*

Serrated polyps include hyperplastic polyps, traditional serrated adenomas and sessile serrated adenomas/polyps. CRC has been observed within serrated lesions (Tanaka et al., 2001) and sessile serrated adenomas progressing to CRC have been reported (Yamauchi et al., 2002, Oono et al., 2009). CRC is thought to develop via serrated polyps in approximately 30% of cases (Bettington et al., 2013). Serrated polyposis is thought to have a hereditary component due to an increased rate of CRC in individuals with serrated polyposis as well as in their FDRs, and a younger age of onset of CRC (Lanspa et al., 2012) but as yet the genetic cause is unidentified.

Mixed polyposis, defined as multiple polyp types and polyps with mixed histologies, has been reported in hereditary mixed polyposis syndrome, which is associated with an increased risk of CRC and in some instances has been linked with duplication of a region upstream of *GREM1,* (Jaeger et al., 2012).

This study included individuals with 20 or more bowel polyps of the various subtypes associated with hereditary CRC. Adenomatous polyps were prioritised, as a separate serrated polyposis study was open for recruitment at the time of this study. Hamartomatous polyposis is more common in childhood, and hence was not prioritised in this study.

### 3.1.2. Referral Guidelines to a Familial Cancer Centre for Individuals with CRC or Adenomatous Polyposis of the Colon

Patients for the discovery cohort were recruited from FCCs across Australia to enrich for individuals with a suspected genetic predisposition to CRC. Individuals with CRC or colonic polyposis are referred to FCCs if the referring clinician suspects that the patient has an underlying genetic predisposition to developing CRC. This is based on a family history of CRC, a young age of onset of CRC, or colonic polyposis suggestive of a hereditary polyposis syndrome, or if a patient requests genetic testing (usually due to a family history of CRC). Australian government endorsed clinical guidelines (https://www.eviq.org.au/Protocol/tabid/66/categoryid/439/id/657/ Referral+ Guidelines+for+Colorectal+Cancer +or+Polyposis+Risk+ Assessment+and+Consideration+of +Genetic+Testing.aspx, Accessed 020317) stipulate the following clinical characteristics in an individual with CRC that warrant a referral to an FCC:

1) CRC diagnosed before 50 years of age in the absence of a family history
2) Multiple Lynch syndrome cancers (CRCs, endometrial, small bowel, gastric, ovarian, pancreatic, transitional cell cancer of the ureter or renal pelvis, cholangiocarcinoma, brain tumour, sebaceous gland tumour, keratoacanthoma) in a single individual
3) A family history of ≥1 FDR or second degree relative (SDR) with CRC or endometrial cancer, with one of the cancers (including the proband) diagnosed prior to age 50 years
4) A family history of ≥2 FDRs or SDRs with a Lynch-syndrome associated cancer, regardless of age of diagnosis

The polyp characteristics that warrant referral to an FCC are:

1)2 or more hamartomatous or juvenile polyps

2) any number of Peutz-Jegher polyps

3) 3 or more adenomatous polyps by 30 years of age, or 20 or more adenomatous polyps at any age

4) 20 or more serrated polyps

While these criteria are intended to select individuals with a high likelihood of having a pathogenic variant in a known CRC-predisposing gene, the criteria also select for the high risk features associated with a genetic predisposition to CRC as described above. As discussed in Chapter 1, there is a proportion of each polyposis syndrome in which a mutation in the corresponding known CRC-predisposing gene is not detected.

### 3.1.3. Aims and Objectives

The aim of the work described in this chapter was to recruit a discovery cohort of individuals with an increased likelihood of harbouring a novel germline predisposition to CRC, for the purpose of germline WES. The specific objectives are to:

a) recruit a high risk CRC discovery cohort, in whom clinical testing for known germline CRC-predisposing mutations has not identified a pathogenic variant.

b) describe the clinical characteristics of the discovery cohort.

## 3.2. Results

### 3.2.1. Inclusion criteria for the Discovery Cohort

Individuals were recruited to this study if they were 18 years of age or older and able to give informed consent and met one or more of the following criteria:

1) Diagnosis of CRC at any age with a family history of CRC in a FDR or SDR

2) Diagnosis of CRC before age 40 years, or

3) Diagnosis of 20 or more bowel polyps at any age

All individuals must have undergone clinical diagnostic testing and had identified no pathogenic variant in a known CRC or polyposis gene identified. CRCs had to be mismatch repair proficient, as determined by the presence of all four mismatch repair proteins by immunohistochemistry (IHC) or by microsatellite testing. An exception was made for individuals with mismatch repair deficient CRC, if germline testing of the relevant mismatch repair gene performed by an FCC did not identify a pathogenic variant or mismatch repair status was discordant between familial samples. For rare cases in which MMR IHC or microsatellite testing had not been performed (usually because the date of diagnosis pre-dated routine MMR testing), germline WES data was checked for pathogenic variants in MMR genes (Section 3.2.4) to ensure that individuals with Lynch syndrome were excluded. Individuals with young onset CRC in the presence of inflammatory bowel disease were excluded from the discovery cohort.

#### 3.2.1.1 *Recruitment of additional affected family members*

Relatives of probands with CRC were invited to participate in this study through the proband. A 'consent to share contact details' form was passed to affected relatives by the proband enrolled in the study. If the affected relative consented to be contacted about the study, the study was explained, and informed consent to participate in the study was obtained.

### 3.2.2. Sites of recruitment

Following ethics approval (Methods Section 2.1.9.1), individuals fulfilling the inclusion criteria were invited to participate in this study. FCCs in the states of Victoria, New South Wales, Queensland, South Australia and Western Australia were invited to participate in this study. FCCs at the following hospitals enrolled patients in this study: PMCC (VIC, Australia), Royal Melbourne Hospital (VIC, Australia), Monash Medical Centre (VIC, Australia), King Edward Memorial Hospital (WA, Australia) and Prince of Wales Hospital (NSW, Australia). FCCs in Queensland and South Australia did not participate due to a delay in ethics approval and competing studies.

Three generation pedigrees of all individuals who met the inclusion criteria were reviewed. Probands with a younger age of onset and/or multiple relatives with CRC were prioritised for WES to increase the likelihood of an underlying genetic predisposition.

Germline DNA from 17 patients who had developed CRC prior to age 40 years was also obtained from the Molecular and Cellular Oncology (MCO) Group study cohort courtesy of Prof Robyn Ward, (University of New South Wales, NSW, Australia) (Ward and Hawkins, 2013, Hawkins et al., 2009). The MCO study prospectively enrolled patients with surgically resected CRC between 1994 and 2010.

### 3.2.3. Clinical Characteristics of the Discovery Cohort

In total, 54 individuals (28 females and 26 males) from 49 families were included in the discovery cohort. Fifty of the 54 individuals in the discovery cohort had developed CRC, while 4 individuals had colonic polyposis but had not developed CRC. The age of diagnosis of CRC or colonic polyposis ranged from 23 years to 66 years, with a median age of 36.5 years. The clinical characteristics, pre-genetic and genetic testing results for all individuals included in the discovery cohort are summarised in Table 3.1A-C.

#### *3.2.3.1. Family history of CRC*

Family history of CRC (defined as at least 1 FDR or SDR affected by CRC) was collected as 3-generation pedigrees. Self-reported cancers within a family were verified by the FCC through the local cancer registries where possible. For 17 individuals (PC13 -PC29) recruited from the MCO study, only self-reported family histories were available. Of the 50 probands with CRC, 29 had a family history of CRC (58%). 20 individuals had one or more FDRs with CRC, while 9 individuals had one or more SDRs with CRC without an affected FDR. The mean age of probands with a family history of CRC was 42 years (range 23 to 66 years).

For 5 of the 29 familial cases, 2 affected individuals from each family agreed to participate in the study. These consisted of 2 affected FDRs from 4 families, and 2 affected SDRs from 1 family. In the remaining familial cases, a single individual from each family was sequenced. More distantly related affected family members were invited to participate, but they did not respond to the invitation.

#### *3.2.3.2. Isolated Young Onset CRC*

There were 20 isolated individuals who developed CRC before the age of 40 years (mean diagnosis age 32 years, range 23-39 years) with no family history of CRC. A further isolated individual was diagnosed with CRC and 30 colonic polyps at age 45 years, and was included in the study due to colonic polyposis.

#### *3.2.3.3. Site of CRC*

Fifty individuals had a diagnosis of 52 CRCs, (47 individuals with adenocarcinomas; 3 individuals with appendiceal mucinous tumours (AMTs). Two individuals had been diagnosed with 2 CRCs. One individual (PC49) had synchronous CRCs in the transverse colon and sigmoid, while another individual (PC6) had metachronous CRCs; a sigmoid cancer at age 39 years and an ascending colon cancer at age 70 years.

One individual had a diagnosis of pseudomyxoma peritonei at the age of 36 years that was presumed to have arisen in an AMT (Ronnett et al., 1997). The pathology report for the appendix was not available for this individual (PC1). The site of CRC for the 50 individuals is

shown in Table 3.1. The number of CRCs at each site is as follows: rectum – 24; rectosigmoid – 2; sigmoid – 14; descending colon – 1; transverse colon – 2; ascending colon – 3; caecum – 2; appendix - 3. For one individual the CRC site was not specified in the pathology report, however the colon specimen containing the CRC consisted of appendix, caecum and ascending colon, hence it was a right sided CRC.

#### *3.2.3.4. Individuals with CRC and ten or more colonic polyps*

Of the 50 individuals with CRC, 7 had CRC and at least 10 colonic polyps at the time of enrolment into the study. The estimated number of colonic polyps and polyp types are summarised in Table 3.1B.

One individual (PC44) had tubular adenomas and tubulovillous adenomas only. Another individual (PC45) had hyperplastic polyps only. Four individuals had more than one polyp type. PC50 had a history of colonic polyposis at age 14 years, for which a colectomy was performed overseas, many years prior to immigrating to Australia. Information on the specific polyp type or estimated number was not available to the FCC, although the polyps were thought to be 'mainly adenomas' according to a clinic note. PC46 and PC47 are siblings.

The polyp descriptions in the pathology reports were not precise enough to determine if the modified World Health Organisation criteria for serrated polyposis syndrome (Bosman et al., 2010) were met. Polyp locations were often not precisely specified, and often only the largest and smallest diameters of the polyps were given.

#### *3.2.3.5. Characteristics of individuals with polyposis without CRC*

Four individuals had colonic polyposis but had not developed CRC at the time of entry into the study (Table 3.1C). PC51 had a family history of FAP, although the genetic cause had not been identified. This patient had colonoscopies in a paediatric hospital up to the age of 18 years and was not thought to have familial adenomatous polyposis at that time. Subsequently, the patient presented at around age 30 with per rectal bleeding and was found to have multiple tubular adenomas in the colon (number not available). The patient underwent a subtotal colectomy. The patient's mother died of CRC at age 44 years and had a clinical diagnosis of FAP and her sister died young from an unspecified cancer. There was no other family history of CRC or colonic polyposis in the family. Testing of the *APC* gene in 2000 by protein truncation test, real time PCR, fluorescence *in situ* hybridisation and Southern blotting did not reveal a mutation. The patient had ongoing screening gastroscopies and colonoscopies. Multiple gastric and duodenal tubular adenomas and tubulovillous adenomas were removed. In 2012, the patient was then found to have at least 20 tubular adenomas in the duodenum on gastroscopy and was referred back to the Familial Cancer Centre. Multiplex Ligation-dependent Probe Amplification (MLPA) of the *APC* gene and full sequencing of the *MUTYH* gene did not identify any mutations.

PC52 underwent an initial colonoscopy at the age of 33 years (indication not given in clinical notes), and was found to have more than 20 polyps in the left colon. A biopsy of one polyp revealed a low-grade tubulovillous adenoma. A repeat colonoscopy identified 5 more polyps in the transverse colon in addition to the left sided polyps. Polypectomies of 10 of these polyps revealed hyperplastic polyps. A third colonoscopy in 6 months identified polyps throughout the colon, with a further 5 hyperplastic polyps retrieved as well as 2 traditional serrated

adenomas. This patient potentially has serrated polyposis syndrome, although based on the available sampled polyp information (17 sampled serrated polyps) the arbitrary WHO criteria for serrated polyposis syndrome have not yet been met. There was no family history of colonic polyposis or CRC in FDR or SDR. Full gene sequencing of the *MUTYH* and *APC* genes, and MLPA of the *APC* gene, did not detect any mutations.

PC53 had a colonoscopy at the age of 65 years following a positive faecal occult blood test performed as part of population screening. At least 30 polyps were seen throughout the colon. Biopsied polyps from all parts of the colon revealed low grade tubular adenomas. There was no family history of polyposis or CRC. Full gene sequencing and MLPA of the *APC* gene, and full gene sequencing of the *MUTYH* gene, did not identify any mutations.

PC54 underwent screening colonoscopies due to a family history of CRC, and had 47 colonic polyps removed between the age of 55 and 72 years. The polyps were tubular adenomas and tubulovillous adenomas, with the exception of 3 hyperplastic polyps. PC54 had 3 FDR (siblings) and 3 SDR with CRC. One of the SDRs is PC41 (enrolled in this study), who developed CRC at the age of 38 years, without colonic polyposis.

#### *3.2.3.6. Pre-genetic screening by IHC of MMR Proteins*

IHC for MMR proteins was performed on CRCs from 42 of the 50 individuals with CRC. Thirty-seven were MMR proficient for MLH1, MSH2, MSH6 and PMS2. For 2 individuals (PC38 and PC49), MMR IHC had been performed on 2 and 3 of the 4 MMR proteins respectively. PC38 was also tested for microsatellite instability and was microsatellite stable. For PC49, IHC for PMS2 had not been performed as the CRC was tested 13 years ago. Importantly, no pathogenic variants in MMR genes were detected in WES data from the germline of PC49 (Chapter 4).

Three individuals (PC7, PC36 and PC40) had MMR deficient CRCs with no evidence for a germline predisposition found on further testing. PC7 (age 40) had absent staining for MSH2 and MSH6, however full sequencing and MLPA of *MSH2* and *MSH6* using blood DNA was uninformative. *EPCAM* was not tested by the FCC, as testing took place prior to routine *EPCAM* testing. PC6 – the sibling of PC7 had developed MMR proficient CRC at age 39, suggesting young onset cancer predisposition with discordant MMR involvement.

PC 36 had equivocal staining of MSH6, however full sequencing and MLPA of *MSH6* did not identify a pathogenic variant. PC40 (age 65) had absent staining of MLH1 and PMS2, however subsequent testing detected a BRAF:c.1779T>A:p.V600E mutation that is associated with somatic hypermethylation of the *MLH1* promoter (Domingo et al., 2004). PC40 had a family history of CRC but this did not meet Amsterdam I criteria.

MMR IHC results were not available for 3 individuals whose CRCs were resected in 1992 (PC31, age 42), 2001 (PC3, age 57) and 2005 (PC11, age 37). Pathogenic variants were not identified in germline WES of these 3 individuals (Chapter 4).

MMR IHC was not performed on CRCs from the following 5 individuals: 2 individuals with AMTs (PC1, PC9), 2 individuals who had colonic polyposis and CRC (PC46, PC50) and 1 individual who had a microsatellite low CRC (PC32, age 41). Inherited pathogenic variants in the mismatch repair genes have not been reported in AMTs (Taggart et al., 2013, Racek et al., 2011).

### *3.2.3.7. Further Diagnostic Genetic Testing by FCCs for Germline Mutations in known CRC-Predisposing Genes*

Testing for inherited pathogenic variants in the *APC* gene was performed on all individuals with polyposis without CRC (Table 3.1C), and on 5 of the 7 individuals with CRC and colonic polyposis (Table 3.1B). PC45 did not have *APC* testing as only hyperplastic polyps had been found, while PC46 was a sibling of PC47 who had already been tested. A further 2 individuals from the cohort had *APC* testing. PC12, an individual with CRC and 5 tubular adenomas, had a family history of colonic polyps (tubulovillous adenomas and hyperplastic polyps). PC12's brother required a hemi-colectomy for colonic polyps (type unknown) at the age of 48 years, while PC12's mother had at least 20 colonic polyps (type unknown) and required a bowel resection at age 67 years. PC19 had 2 tubular adenomas and 1 hyperplastic polyp at the time of diagnosis of a rectal cancer, and a SDR with a history of 'severe polyps'. Testing for *APC* pathogenic variants was performed by full gene sequencing and MLPA in all individuals apart from PC51. PC51 had a protein truncation test, fluorescence in situ hybridisation, and subsequently real time PCR and MLPA, as these were the available technologies at the time of testing. All *APC* genetic testing by the FCCs was uninformative.

Testing for homozygous or compound heterozygous pathogenic variants in the *MUTYH* gene was performed on all individuals who had *APC* testing. *MUTYH* testing was also performed in 4 individuals with isolated young onset CRC (PC4, PC10, PC30, PC37), 1 individual with young onset CRC (PC6) with an affected sibling (PC7), and 1 individual (PC40) with CRC and 6 polyps, some of which were hyperplastic. PC40 also had a family history of CRC in 3 siblings. Individuals from the MCO study were included in this study on the basis that no pathogenic variants had been found through testing at an FCC, however the specific genes tested for were not provided. There were no colonic adenomas found in any of the MCO study individuals. All *MUTYH* testing by the FCCs was uninformative. *MUTYH* was not tested in all isolated young onset patients with CRC, possibly because routine testing had not been introduced into the clinic at the time the patients were evaluated by the FCCs. Genetic testing for *POLE* and *POLD1* had not yet been introduced to the FCCs at the time of testing of individuals recruited to this study.

## 3.3. Discussion

The aim of the work presented in this chapter was to recruit individuals suspected of having a germline predisposition to CRC, but in whom no CRC-predisposing mutation had been found through diagnostic testing in an FCC. The level of testing performed was dependent on the time of FCC referral and the clinical testing available at that time. This cohort formed the discovery cohort for identifying candidate CRC-predisposing genes. The 54 recruited individuals represent patients whose susceptibility to CRC remained unexplained after pre-genetic and genetic investigations performed by the FCC.

### 3.3.1. Strengths and limitations of the discovery cohort

This discovery cohort is derived from the Australian population, which has not been widely reported on. Only one germline WES study of 16 families with CRC, recruited from the Colon Cancer Family Registry (that includes families from the United States, Canada and Australia) has been reported (Derycke et al., 2013).

Forty-two percent of recruited individuals with CRC (n=21) did not have an affected FDR or SDR. These cases developed CRC before age 40 years (except for one individual who was diagnosed with CRC and 30 colonic polyps at age 45 years). This may reflect referral patterns, where patients with young-onset CRC or colonic polyposis may be more readily referred to FCCs than individuals with a family history of CRC unless aberrant MMR IHC is found in the latter.

There is heterogeneity in the clinical phenotypes included in the discovery cohort. This suggests that the unexplained suspected familial cases seen in the Australian FCCs that participated in the study may have distinct underlying pathogenic variants and reflect genetic subtypes of familial CRC. The discovery cohort included a rare CRC phenotype, a family of 2 individuals, and a 3$^{rd}$ unrelated individual with AMTs. AMTs occur in 0.12 per million people per year (Jemal et al., 2009). Two familial cases have been described in the literature: a pair of affected identical twin brothers in their 30s (Shih et al., 2001), and a pair of siblings aged 69 and 77 years (Racek et al., 2011). A genetic predisposition to AMTs has not been identified to date, however assessment for germline predisposition has been limited in these 2 families. LOH of *APC* was assessed in the affected twins, and was found in the AMT of one twin but not the other. In the second reported familial case, sequencing of *5* genes, *MLH1, MSH2, MSH6, PKD1* and *PKD2* (the latter two genes were tested due the finding of renal cysts in one of the affected individuals) was performed in the 69 year old sibling, however no pathogenic variants were identified. The inclusion of 3 individuals with AMT in our discovery cohort allows germline WES to be performed for this rare disease.

Estimated polyp count is limited by the extent to which the colon is assessed either by colonoscopy or surgical resection, availability of reports from such procedures, and technical factors such as poor bowel preparation. Assessment of polyp histopathology may also be influenced by the number of polyps that were sampled. Hence individuals thought to have a single histopathologic subtype of colonic polyp may in fact have other types that were not sampled. The significance of different histopathologic subtypes is still being elucidated, however it likely reflects different genetic pathways of tumorigenesis. Underestimation of polyp number or under-reporting of polyp type (such as serrated polyps) limits the number of

patients who meet the referral criteria to FCCs. Despite the modified WHO criteria for serrated polyposis syndrome (Bosman et al., 2010) predating most of the pathology reports, the diagnostic criteria does not appear to be taken into consideration when reporting. While the criteria are arbitrary due to the lack of a genotypic correlate at present, the lack of classification hampers research attempts to study this syndrome.

A limitation with recruiting individuals referred to FCCs is the dependence on the primary care provider or medical specialist for referral. An increased risk of a familial predisposition to CRC may not be recognised by the primary treating doctor, hence there may be individuals with an unexplained genetic predisposition to CRC who are not referred to an FCC for further evaluation. Other factors include a lack of awareness of the utility of germline genetic testing, and possibly the challenges of travelling to a FCC for patients who live in very remote regions of Australia.

### 3.3.2. Challenges in recruitment

More distally related relatives (eg third degree relatives) of probands who developed CRC in several cases could not be recruited into the study as they lived overseas, or did not respond to the invitation to participate. This may reduce the likelihood of identifying a CRC-predisposing gene in the family (assuming a germline predisposition to CRC was present), as more closely related relatives share a greater proportion of their genome, thereby increasing the number of shared private variants that may be difficult to distinguish from CRC-predisposing genetic variants.

### 3.3.3. Composition of the Discovery cohort

The preponderance of CRCs distal to the splenic flexure seen in our discovery cohort is consistent with other studies of young onset CRC (Dozois et al., 2008, You et al., 2012). Seventy-seven percent of the CRCs in our cohort were distal to the splenic flexure, comparable to 72% of CRCs diagnosed before age 50 years using data from the New South Wales Central Cancer Registry (Boyce et al., 2016).

Several distinct histopathological subgroups were identified in our discovery cohort. This includes individuals with colonic polyposis and individuals with AMTs.

### 3.3.4. Summary of the Discovery Cohort

Fifty-four individuals were recruited into the discovery cohort. Sixty-eight percent had CRC or colonic polyposis before age 40 years, and 58% have an affected FDR or SDR. The clinical characteristics of the discovery cohort are heterogeneous, which may decrease the likelihood of identifying a common novel CRC-predisposing gene if this clinical heterogeneity reflects underlying genetic heterogeneity, as WES studies have had the most success in monogenic diseases. Clinical heterogeneity does not rule out a common genetic predisposition however. For example, pathogenic variants in *MUTYH* have demonstrated variable clinical phenotypes (Sieber et al., 2003, Guarinos et al., 2014, Farrington et al., 2005). The discovery cohort

assembled here is a representation of the cases that are seen at the FCCs due to a suspected underlying germline predisposition to CRC.

Germline variants identified in WES data from the 54 individuals in this discovery cohort are analysed in detail in Chapter 4.

**Table 3-1. Clinical, pre-genetic and genetic test results for individuals recruited to the discovery cohort. A) Individuals with CRC. B) Individuals with CRC and ≥10 colonic polyps. C) Individuals with colonic polyposis without CRC.**

## A) Individuals with CRC

| Sample ID | Age | Reason for Inclusion | Number of FDR with CRC | Number of SDR with CRC | CRC location | Polyp Type(s) | Estimated Cumulative Polyp Count at Recruitment | MLH1 IHC | MSH2 IHC | MSH6 IHC | PMS2 IHC | FCC Reason for Germline Testing | Gene | Testing Method | Result |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **PC1** | 36 | Young Onset | 0 | 0 | presumed appendix | - | - | - | - | - | - | - | - | - | - |
| **PC2*** | 38 | Young Onset/Familial | 1 | 0 | rectum | - | - | P | P | P | P | - | - | - | - |
| **PC3*** | 57 | Familial | 1 | 0 | rectum | 1.TVA 2.mixed (TVA+ conventional serrated adenoma) 3.TA | 3 | - | - | - | - | - | - | - | - |
| **PC4** | 28 | Young Onset | 0 | 0 | sigmoid | 1.villous adenoma 2.hyperplastic | ≥3 | P | P | P | P | hyperplastic polyp | *MUTYH* | PCR amplification of exons 7, 12 and 13 | negative |
| **PC5** | 41 | Familial | 2 | 1 | rectum | 1.TVA | 1 | P | P | P | P | - | - | - | - |
| **PC6*** | 39 | Young Onset/Familial | 1 | 0 | sigmoid | 1. villous adenoma | 1 | P | P | P | P | recessive | *MUTYH* | Single Nucleotide Primer extension analysis for 2 European founder mutations | negative |
| **PC7*** | 40 | Familial | 1 | 0 | ascending colon | 1.TA | 1 | P | D | D | P | IHC | *MSH2, MSH6* | full gene sequencing, MLPA | negative |
| **PC8*** | 51 | Familial | 1 | 0 | appendix | - | - | P | P | P | P | - | - | - | - |
| **PC9*** | 66 | Familial | 1 | 0 | appendix | - | - | - | - | - | - | - | - | - | - |
| **PC10** | 23 | Young Onset | 0 | 0 | rectum | | | P | P | P | P | isolated young onset | *MUTYH* | full gene sequencing | negative |
| **PC11** | 37 | Young Onset | 0 | 0 | rectum | - | - | - | - | - | - | - | - | - | - |

**FDR, first degree relative. SDR, second degree relative. CRC, colorectal cancer. IHC, immunohistochemistry. FCC, Familial Cancer Centre. TVA, tubulovillous adenoma. TA, tubular adenoma. P, proficient. D, deficient. H, heterogeneous. PCR, polymerase chain reaction. MLPA, multiplex ligation probe amplification. D-HPLC, denaturing high performance liquid chromatography. N/A, not available.**

*denotes related individuals (PC2, PC3); (PC6, PC7); (PC8, PC9); (PC46, PC47); (PC41, PC54)

A) Individuals with CRC, continued

| Sample ID | Age | Reason for Inclusion | Number of FDR with CRC | Number of SDR with CRC | CRC location | Polyp Type(s) | Estimated Cumulative Polyp Count at Recruitment | MLH1 IHC | MSH2 IHC | MSH6 IHC | PMS2 IHC | FCC Reason for Germline Testing | Gene | Testing Method | Result |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **PC12** | 55 | Familial | 1 | 3 | sigmoid | 1.TVA 2.TA 3.mixed TA and hyperplastic | 5 | P | P | P | P | polyps | *APC, MUTYH* | full gene sequencing | negative |
| **PC13** | 35 | Young Onset | 0 | 0 | sigmoid | - | - | P | P | P | P | - | - | - | - |
| **PC14** | 39 | Young Onset | 0 | 0 | sigmoid | - | - | P | P | P | P | - | - | - | - |
| **PC15** | 31 | Young Onset/Familial | 1 | 0 | sigmoid | - | - | P | P | P | P | - | - | - | - |
| **PC16** | 23 | Young Onset/Familial | 1 | 0 | sigmoid | - | - | P | P | P | P | - | - | - | - |
| **PC17** | 34 | Young Onset | 0 | 0 | rectum | - | - | P | P | P | P | - | - | - | - |
| **PC18** | 39 | Young Onset/Familial | 1 | 0 | rectum | - | - | P | P | P | P | - | - | - | - |
| **PC19** | 39 | Young Onset | 0 | 0 | rectum | 1.Adenoma | 1 | P | P | P | P | - | - | - | - |
| **PC20** | 37 | Young Onset | 0 | 0 | rectum | - | - | P | P | P | P | - | - | - | - |
| **PC21** | 38 | Young Onset | 0 | 0 | rectum | - | - | P | P | P | P | isolated young onset | *MUTYH* | full gene sequencing | negative |
| **PC22** | 32 | Young Onset | 0 | 0 | rectum | - | - | P | P | P | P | - | - | - | - |
| **PC23** | 32 | Young Onset | 0 | 0 | rectum | 1.Adenoma | 1 | P | P | P | P | - | - | - | - |
| **PC24** | 24 | Young Onset | 0 | 0 | rectum | - | - | P | P | P | P | - | - | - | - |
| **PC25** | 29 | Young Onset/Familial | 1 | 0 | rectum | - | - | P | P | P | P | - | - | - | - |
| **PC26** | 34 | Young Onset | 0 | 0 | rectum | | - | P | P | P | P | - | - | - | - |
| **PC27** | 34 | Young Onset | 0 | 0 | sigmoid | 1.Adenoma | 1 | P | P | P | P | - | - | - | - |
| **PC28** | 28 | Young Onset/Familial | 1 | 0 | sigmoid | - | - | P | P | P | P | - | - | - | - |
| **PC29** | 27 | Young Onset | 0 | 0 | ascending colon | 1.Adenoma | 1 | P | P | P | P | - | - | - | - |
| **PC30** | 34 | Young Onset/Familial | 0 | 1 | rectum | 1.Hyperplastic 2.TA | 3 | P | P | P | P | polyps | *APC, MUTYH* | N/A | negative |

FDR, first degree relative. SDR, second degree relative. CRC, colorectal cancer. IHC, immunohistochemistry. FCC, Familial Cancer Centre. TVA, tubulovillous adenoma. TA, tubular adenoma. P, proficient. D, deficient. H, heterogeneous. PCR, polymerase chain reaction. MLPA, multiplex ligation probe amplification. D-HPLC, denaturing high performance liquid chromatography. N/A, not available.

A) Individuals with CRC, Continued

| Sample ID | Age | Reason for Inclusion | Number of FDR with CRC | Number of SDR with CRC | CRC location | Polyp Type(s) | Estimated Cumulative Polyp Count at Recruitment | MLH1 IHC | MSH2 IHC | MSH6 IHC | PMS2 IHC | FCC Reason for Germline Testing | Gene | Testing Method | Result |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **PC31** | 42 | Familial | 0 | 3 | sigmoid | - | - | - | - | - | - | - | - | - | - |
| **PC32** | 41 | Familial | 1 | 0 | transverse colon | - | - | # | # | # | # | - | - | - | - |
| **PC33** | 31 | Young Onset/Familial | 0 | 2 | rectum | - | - | P | P | P | P | - | - | - | - |
| **PC34** | 25 | Young Onset/Familial | 0 | 1 | sigmoid | 1.hyperplastic | 1 | P | P | P | P | - | - | - | - |
| **PC35** | 24 | Young Onset/Familial | 0 | 1 | sigmoid | - | - | P | P | P | P | - | - | - | - |
| **PC36** | 37 | Young Onset/Familial | 0 | 1 | rectum | 1.TVA | 1 | - | - | H | P | IHC | *MSH6* | full gene sequencing, MLPA | negative |
| **PC37** | 27 | Young Onset | 0 | 0 | rectum | 1.hyperplastic | 1 | P | P | P | P | isolated young onset | *MUTYH* | full gene sequencing, MLPA | negative |
| **PC38** | 44 | Familial | 1 | 4 | rectum | 1.hyperplastic 2.TA | 2 | P | P | - | - | young-onset | *MLH1* *MSH2* | D-HPLC | negative |
| **PC39** | 35 | Young Onset/Familial | 0 | 1 | rectum | 1.TVA | 1 | P | P | P | P | - | - | - | - |
| **PC40** | 65 | Familial | 4 | 1 | caecum | 1.TA 2.hyperplastic 3.serrated adenoma | 6 | D$ | P | P | D | polyps | *MUTYH* | full gene sequencing | negative |
| **PC41*** | 38 | Young Onset/Familial | 1 | 2 | caecum | - | - | P | P | P | P | - | - | - | - |
| **PC42** | 32 | Young Onset | 0 | 0 | rectum | 1.TVA 2.hyperplastic | 3 | P | P | P | P | - | - | - | - |
| **PC43** | 31 | Young Onset | 0 | 0 | sigmoid | - | - | P | P | P | P | - | - | - | - |

FDR, first degree relative. SDR, second degree relative. CRC, colorectal cancer. IHC, immunohistochemistry. FCC, Familial Cancer Centre. TVA, tubulovillous adenoma. TA, tubular adenoma. P, proficient. D, deficient. H, heterogeneous. PCR, polymerase chain reaction. MLPA, multiplex ligation probe amplification. D-HPLC, denaturing high performance liquid chromatography. N/A, not available.

*denotes related individuals (PC2, PC3); (PC6, PC7); (PC8, PC9); (PC46, PC47); (PC41, PC54)
#PC32:MSI-low (1 of 6 microsatellite markers; DS123 stable; Bat25 unstable; Bat26 stable; D5S346 stable; D17S250 stable; Bat40 stable)
$BRAF mutation positive

B) Individuals with CRC and ≥10 colonic polyps

| Sample ID | Age | Reason for Inclusion | No. of FDR with CRC | Number of SDR with CRC | CRC location | Polyp Type | Estimated Cumulative Polyp Count at Recruitment | MLH1 IHC | MSH2 IHC | MSH6 IHC | PMS2 IHC | FCC Reason for Germline Testing | Gene | Testing Method | Result |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **PC44** | 62 | Familial | 2 | 0 | descending colon | 1.TVA 2.TA | 21 | P | P | P | P | polyposis | *APC MUTYH* | full gene sequencing; MLPA (APC only) | negative |
| **PC45** | 26 | Young Onset | 0 | 0 | R sided | 1.hyperplastic | 10 | P | P | P | P | hyperplastic polyps | *MUTYH* | full gene sequencing | negative |
| **PC46*** | 63 | Polyposis/Familial | 1 | 0 | Recto-sigmoid | 1.TVA 2.TA 3.hyperplastic | 25 | - | - | - | - | (see PC47) | - | - | - |
| **PC47*** | 55 | Polyposis/Familial | 1 | 0 | rectum | 1.TVA 2.TA 3.sesile serrated polyp 4.hyperplastic | 80 | P | P | P | P | polyposis | *APC MUTYH* | N/A | negative |
| **PC48** | 45 | Polyposis | 0 | 0 | Recto-sigmoid | 1.TVA 2.TA 3.sesile serrated adenoma 4.hyperplastic 5.mixed TA and serrated adenoma | 30 | P | P | P | P | polyposis | *APC MUTYH* | - | - |
| **PC49** | 30 | Young Onset/Familial | 0 | 2 | transverse colon; sigmoid (synchro-nous) | 1.Adenoma 2.hyperplastic | 40 | P | P | P | - | polyposis | *APC* | - | - |
| **PC50** | 48 | Polyposis/Familial | 0 | 1 | rectum | unknown (resected overseas) | N/A | - | - | - | - | polyposis | *APC MUTYH* | full gene sequencing, MLPA | negative |

FDR, first degree relative. SDR, second degree relative. CRC, colorectal cancer. IHC, immunohistochemistry. FCC, Familial Cancer Centre. TVA, tubulovillous adenoma. TA, tubular adenoma. P, proficient. D, deficient. MLPA, multiplex ligation probe amplification. N/A, not available.

C)Individuals with colonic polyposis without CRC

| Sample ID | Age | Reason for Inclusion | No. of FDR with CRC | Number of SDR with CRC | CRC location | Polyp Type | Estimated Cumulative Polyp Count at Recruitment | MLH1 IHC | MSH2 IHC | MSH6 IHC | PMS2 IHC | FCC Reason for Germline Testing | Gene | Testing Method | Result |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **PC51** | 30 | Polyposis | 1 | 1 | - | N/A | N/A | - | - | - | - | polyposis | *APC* *MUTYH* | *APC*-PTT, FISH, real time PCR, MLPA; *MUTYH*-full gene sequencing | negative |
| **PC52** | 33 | Polyposis | 0 | 0 | - | 1.hyperplastic 2.serrated adenoma 3.TA | 20 | - | - | - | - | polyposis | *APC* *MUTYH* | full gene sequencing; MLPA (*APC* only) | negative |
| **PC53** | 64 | Polyposis | 0 | 0 | - | 1.TA | 30 | - | - | - | - | polyposis | *APC* | full gene sequencing | negative |
| **PC54*** | 55 | Familial/ Polyposis | 3 | 3 | - | 1.TVA 2.TA 3.hyperplastic | 47 | - | - | - | - | polyposis | *APC* *MUTYH* | N/A | negative |

FDR, first degree relative. SDR, second degree relative. CRC, colorectal cancer. IHC, immunohistochemistry. FCC, Familial Cancer Centre. TVA, tubulovillous adenoma. TA, tubular adenoma. PTT, Protein Trunctaion Test. PCR, polymerase chain reaction. MLPA, multiplex ligation probe amplification. N/A, not available.

# Chapter 4. Identifying Candidate Familial CRC-predisposing Genes in WES data from the Discovery Cohort

## 4.1. Introduction

Chapter 3 described the creation of a discovery cohort of 54 individuals with a high likelihood of having an inherited monogenic predisposition to CRC. Germline WES and variant calling was performed in each of these individual as described in the Methods (Chapter 2 sections 2.2.4 and 2.2.5). This chapter focuses on analysing the data to identify possible novel candidate genes predisposing to CRC.

Identification of pathogenic variants in known CRC-predisposing genes through WES that were missed or not tested through the FCC, is also discussed but data from these individuals were then excluded from further analysis. Candidate CRC-predisposing genes identified from the discovery cohort were screened in a larger validation cohort described in Chapter 5.

While GWAS studies have identified common low penetrance variants that may account for a proportion of inherited CRCs, it is possible that the 'missing heritability' for CRC is accounted for by the inheritance of rare moderate/high penetrance disease-causing variants in as yet unidentified genes. The vast majority of disease-causing mutations for Mendelian disease have been identified in the coding regions or essential splice sites of a gene. The following assumptions were made regarding the germline predisposition to CRC in the discovery cohort: 1) a pathogenic variant predisposing to CRC was present in protein-coding regions or essential splice sites within the genome, 2) the commonest mechanism for disease is loss of gene function, rather than activating gain of function mutations, 3) mutations in the gene confers a moderate to high risk of developing CRC and the pathogenic variant will be rarely seen or absent in the unaffected population, and 4)the pathogenic variant follows Mendelian inheritance.

The approach to identifying candidate genes predisposing to CRC was divided into 2 stages:

A) Identifying rare germline variants predicted to have a functional impact on the protein *i.e.* potentially pathogenic variants (PPVs)

B) Prioritising candidate CRC-predisposing genes with these PPVs

The following sections describe the approach used to identify PPVs from WES data. Once PPVs were identified, several strategies were then used to prioritise candidate CRC-predisposing genes containing these variants, as PPVs in themselves may not predispose to CRC. The strategies used to identify candidate CRC-predisposing genes are described in Section 4.1.2.

### 4.1.1. Identifying PPVs

#### *4.1.1.1. Identification of Rare Variants*

Several publicly available datasets comprising large sequencing studies such as the 1000 Genomes Project (Genomes Project et al., 2012), Exome Sequencing Project (ESP) (Exome Variant Server, n.d.) and Exome Aggregation Consortium (Lek et al., 2016c) can be used to aid in separating rare and potentially pathogenic variants from variants seen in the databases at too high a frequency than is plausible if they were pathogenic. Variants present at ≥0.5%in the general population were excluded. A limitation in these datasets is the lack of phenotypic information.

The 1000 Genomes Project sequenced the genomes of 1092 individuals randomly selected from 14 populations, using a combination of low-coverage whole genome sequencing, WES and dense SNP genotyping. The ESP comprises the exomes of 6503 unrelated individuals (4300 European Americans and 2203 African Americans) from either healthy control populations or populations with early-onset cardiovascular or lung diseases. A much larger dataset, ExAC became available midway through this project. ExAC combined 14 population or disease-based exome sequencing studies, by performing standardised variant calling on raw sequencing data from 60,706 study individuals. The disease cohorts comprised metabolic, cardiac, psychiatric and inflammatory bowel disorders, as well as individuals diagnosed with cancer from TCGA. As individuals with a germline predisposition to cancer may be present in the TCGA cohort, the subset of the ExAC dataset that excluded 7601 TCGA samples was used in the study described in this thesis. This cohort is henceforth referred to as the Non-TCGA Subset of ExAC.

For the purpose of this study, a rare allele is defined as an allele that is observed in less than 1% of the population (Bamshad et al., 2011). The optimal allele frequency cut-off for identifying rare disease-causing alleles is unknown but is guided by the overarching rule that a highly penetrant allele cannot be present in the population at a significantly higher frequency than the disorder it causes. Determining an allele frequency cut-off for CRC-predisposing variants inherited in an autosomal recessive manner is even more challenging, as the heterozygous state may not result in an increased CRC risk. While the literature suggests that the population frequency of *MUTYH* heterozygous pathogenic variant carriers is between 1-2% (Cleary et al., 2009), the two most common *MUTYH* mutations, NM_001128425.1:c.536A>G, p.Tyr179Cys and NM_001128425.1:c.1187G>A, p.Gly396Asp, are seen in approximately 0.2-0.4% of European individuals in ExAC. It may be that the larger size of the ExAC dataset gives a more accurate estimate. These two *MUTYH* founder mutations are also specific to the European population, and such population-specific information is not known for novel variants.

#### *4.1.1.2. Rare variants identified through WES – real or artefact?*

Errors in variant calling can confound rare variant identification from WES data. Technical parameters are often used in an attempt to differentiate these errors from true variants.

The GATK variant quality score (QUAL) gives an estimate of a variant caller's accuracy in identifying a true variant. An increase in this logarithmic score gives an exponential increase in confidence that a variant is genuine. For example, a QUAL score of 20 indicates a 99% confidence in the variant call, while a QUAL score of 30 indicates a 99.9% confidence in the

variant call. Sequencing parameters such as read depth, variant frequency and read bidirectionality can also be used to support a variant call. The read depth refers to the number of reads aligned to a particular position in the genome. The alternate read depth is the number of reads that support a variant that differs from the reference genome at that particular location. The variant frequency is the proportion of total reads at a particular position that support the alternate allele. Variant frequencies that are much lower than the diploid state expected of a heterozygous variant (i.e. 0.5) are more likely to represent alignment or sequencing errors than a true variant. If a variant is identified in both forward and reverse reads *i.e.* bidirectional reads, the variant is less likely to have arisen due to an alignment error, as sequencing has occurred in both directions and aligned to the reference genome.

#### *4.1.1.3. Transcript selection and Variant Prioritisation*

The predicted functional consequence of a genetic variant identified through WES can differ depending on the transcript used to make the variant call. A variant may be in a coding region present in one transcript but that region and therefore the variant might be absent in another transcript. Different transcripts may also have different reading frames. WES studies vary in their methods of transcript selection, with some choosing the canonical transcript while others select the transcript in which the variant is predicted to have the most deleterious consequence. It is unclear which approach is better as the exact transcripts that are expressed in the colon and therefore are of the most functional relevance are not known.

Potentially deleterious variants are predicted to affect the function of the encoded protein. Such variants would encompass loss-of-function (LoF) variants (nonsense, frameshift, essential splice site i.e. within 2 base pairs of the intron-exon boundary). LoF variants are generally assumed to be deleterious as these variants are expected to result in a dysfunctional truncated protein, or activate nonsense mediated decay of the RNA transcript. A proportion of missense variants may also cause loss of function through mRNA or protein instability, but without a functional assay these are harder to determine.

##### 4.1.1.3.1. In silico missense classifiers

Differentiating missense variants that are potentially pathogenic from benign ones is a challenge since most amino acid substitutions may not necessarily impact protein function. Performing biological assessment of pathogenicity on large numbers of missense variants is impractical and requires knowledge of gene function. Therefore, a gamut of in-silico tools have been developed to try and predict the impact of missense variants on protein function. Such tools evaluate various aspects of the missense variant, including its predicted change in amino acid biochemistry, nucleotide sequence conservation between species, and functional annotation within protein structure. Examples of commonly used in silico tools include Sorting Intolerant From Tolerant (SIFT) (Ng and Henikoff, 2003), Polymorphism Phenotyping (PolyPhen) (Ramensky et al., 2002), Genomic Evolutionary Rate Profiling (GERP) (Cooper et al., 2005), as well as optimised combination tools such as Consensus Deleteriousness score (Condel) (González-Pérez and López-Bigas, 2011) and Combined Annotation Dependent Depletion (CADD) (Kircher et al., 2014).

CADD is one of the most commonly used tools and utilises 63 different variant annotations and their interactions to differentiate variants that are fixed or nearly fixed in the human genome

from simulated *de novo* variants. The latter are assumed to be selected against and therefore more likely to be deleterious. The variant annotations used by CADD include conservation metrics, functional genomics, in silico predictions of variant impact on protein structure and function, and gene expression. CADD scores are then calculated for all possible SNVs in the human genome, as well as for small indels, using machine learning. High CADD scores indicate that the variant has characteristics associated with simulated de novo variants rather than those seen with fixed/nearly fixed variants. The CADD scores are then ordered from least to most deleterious in a Phred-like scale to give CADD scaled scores. A CADD scaled score of 10 or more corresponds to the top ten percent most deleterious variants according to the CADD model, while a CADD score of 20 or more corresponds to the top one percent of deleterious variants, and so on.

CADD scores were compared to an expertly curated classification based on clinical and genomic data, using variants in mismatch repair genes that had been classified by the Variant Interpretation Committee of the International Society for Gastrointestinal Hereditary Tumours (InSiGHT) (van der Velde et al., 2015). Using a cumulative link model, this paper found a strong positive correlation between CADD scores and InSiGHT classification. False positives did occur, but these could be compensated for by using a population allele frequency cut-off to exclude common variants. False negatives tended to occur with loss-of-function variants as CADD was designed as a quantitative assessment of SNVs rather than structural variants.

### 4.1.2. Approaches to identifying candidate CRC-predisposing genes harbouring potentially pathogenic variants

Once potentially pathogenic variants have been identified, the genes that harbour these variants are then prioritised as potential CRC-predisposing genes using several approaches.

#### *4.1.2.1. Agnostic Approach*

The agnostic approach relies on candidate CRC-predisposing genes being recurrently mutated within the discovery cohort at a frequency higher than expected based on publicly available control cohorts. This approach is applied in two different settings. Firstly, an inter-family analysis looks at the frequency across all the families. Genes that harbour PPVs in multiple unrelated individuals from the discovery cohort are prioritised as candidate CRC-predisposing genes. This approach is further refined using an intra-family analysis by prioritising variants that are carried by multiple affected members of the same family.

#### *4.1.2.2. Candidate Gene Approach*

Amongst 54 exomes, it is possible that a gene predisposing to CRC is only mutated in one of the study subjects due to disease heterogeneity, with only one affected individual in a family being available for WES. Such genes would not be identified by the analysis methods described in the previous section, which relies on the observation of a recurrently mutated gene in the discovery cohort. As many of the known CRC-predisposing genes are within well-characterised pathways (Table 4.1), it is possible that a mutation in a different gene within the same pathway may also predispose to CRC. Such a candidate gene approach is another approach to identify candidate CRC-predisposing genes. For example, DNA repair, which comprises several distinct pathways including the mismatch repair, base excision and double strand break repair

pathways, is disrupted in Lynch syndrome, *MUTYH*-associated polyposis and polymerase proofreading associated polyposis (Hicks et al., 2010, Pavlov et al., 2006). The bone morphogenetic protein (BMP) pathway is perturbed in the juvenile polyposis syndromes through mutations in *SMAD4* and *BMPR1A*. *APC* mutations give rise to familial adenomatous polyposis through activation of the Wnt pathway (Segditsas and Tomlinson, 2006). The Wnt pathway is also dysregulated in up to 90% of sporadic CRCs (White et al., 2012). In CRCs with wildtype *APC*, activating somatic mutations in *CTNNB1*, a key gene in the canonical Wnt pathway, were found to be important for tumorigenesis (Morin et al., 1997). The Notch signalling pathway is an intercellular signalling process that is important in embryonic development (Artavanis-Tsakonas et al., 1999), and is thought to be perturbed in cancer (Ranganathan et al., 2011). Notch signalling is oncogenic in colorectal mucosa due to inhibition of goblet cell differentiation (Katoh and Katoh, 2007).

**Table 4-1. Pathways associated with known CRC-predisposing genes**

| Pathway | Known CRC-Predisposing Genes |
|---|---|
| DNA repair | *MLH1, MSH2, MSH6, PMS2, MUTYH, POLE, POLD1* |
| BMP | *SMAD4, BMPR1A* |
| Wnt | *APC* |

DNA, deoxyribonucleic acid. BMP, bone morphogenetic protein.

#### *4.1.2.3. Gene Constraint Scores*

Another approach to prioritising candidate CRC-predisposing genes is to use population databases to identify genes that have fewer PPVs per base than expected. Genes with less variation than expected may be evolutionally constrained due to negative selection pressure, a hallmark of dominant disease-associated genes. Examples of gene constraint scores are Residual Variation Intolerance Score (RVIS) (Petrovski et al., 2013) and ExAC gene constraint scores (Lek et al., 2016c). Potentially pathogenic variants within constrained genes would therefore be considered interesting and worthy of further evaluation.

RVIS ranks genes according to their intolerance of common (minor allele frequency >0.1%) functional variation, and is derived from SNVs observed in exomes from the ESP. RVIS takes both gene size and total gene mutation rate into account. Common missense and LoF SNVs from each of the 16956 assessed genes were regressed on the total protein-coding variation seen in each gene, then divided by the estimate of its standard deviation to give the RVIS. Corresponding RVIS percentiles aid in ranking each gene, with the lowest percentiles indicating genes that are intolerant of common functional variation. The RVIS was validated using groups of Mendelian disease-causing genes (annotated in Online Mendelian Inheritance in Man) according to the allelic context in which they were pathogenic (haploinsufficient, dominant negative, de novo, recessive), as well as a group of genes that were considered non-disease causing when mutated. Mendelian disease-causing genes, particularly mutated genes that caused disease when haploinsufficient or that had de novo mutations, had lower RVIS scores than non-disease-causing genes when assessed by logistic regression.

The ExAC gene constraint scores evaluated rare (minor allele frequency <0.1%) variants using the EXAC dataset, which is an order of magnitude bigger than the ESP. For each variant class (i.e. LoF, missense or synonymous variants) the observed number of rare unique variants per gene was compared to an expected number of variants based on stochastic mutation. The expected number of mutations in a gene was determined using a model (Samocha et al., 2014) validated on synonymous variants, which are not expected to be under selection pressure. A z-score indicating deviation from the expected number of variants is given for missense variants, with a score of >3.09 indicating significant constraint of missense variants. As larger genes have an increased probability of acquiring a LoF variant by chance, Z-scores for LoF variants are further categorised into probability of being loss-of function intolerant (pL I≥0.9) or loss-of function tolerant (pLI ≤0.1) using an expectation-maximisation algorithm.

### 4.1.3. Aims and Objectives

The aim of the work described in this chapter is to identify candidate genes predisposing to familial CRC, using WES data obtained from the discovery cohort.

The objectives of this chapter are as follows:

a) identify rare PPVs in WES data using a series of technical and biological filters

b) identify candidate CRC-predisposing genes harbouring PPVs by performing intra-family, inter-family, candidate gene and gene constraint analyses

The top candidate genes from the analyses performed in this chapter were re-sequenced in a validation cohort using a targeted gene panel (Chapter 5).

## 4.2. Methods

### 4.2.1. Examination of variants in Known Colorectal Cancer Predisposing Genes identified by WES

Variants in 11 known CRC-predisposing genes (*APC, MUTYH, POLE, POLD1, MLH1, MSH2, MSH6, PMS2, STK11, SMAD4, BMPR1A)* were examined in the 54 sequenced germline exomes, in order to exclude any exomes with a pathogenic variant in a known CRC-predisposing gene. Such pathogenic variants may not have been detected through the testing methods employed by the FCCs, and would confound subsequent analyses.

The RefSeq gene transcript used in the original paper that identified the gene as CRC-predisposing (Sieber et al., 2003, Palles et al., 2013, Ligtenberg et al., 2009, Howe et al., 2001a) was selected for further analysis. For genes in which a transcript was not identified in the literature, the canonical transcript was selected. Variant annotation is given for the corresponding Ensembl transcript as the output from Variant Effect Predictor.

For genes with an autosomal dominant mode of inheritance, variants were excluded if they occurred at a frequency of 0.5% or greater in any of 3 large publicly available, non-cancer ascertained, germline WES datasets: 1) 1000 Genomes Project (Genomes Project et al., 2012), 2) the ESP (Exome Variant Server, n.d.), or 3) the Non-TCGA Subset of the Exome Aggregation Consortium (ExAC) (Lek et al., 2016c). For genes with an autosomal recessive mode of inheritance (*MUTYH*), variants were excluded if they occurred at a frequency of 5% or greater in the 1000 Genomes Project, the ESP, or the Non-TCGA Subset of ExAC. Variants with an observed allele frequency of 1.0 were considered homozygous. Loss-of-function (LoF) variants, namely nonsense, frameshift and essential splice site variants, as well as in-frame indels and missense variants, were selected for further analysis. For in-frame indels and missense variants, only those with a CADD scaled score of 10 or more were considered. ClinVar (Landrum et al., 2016) annotations were considered for variants of uncertain significance. Samples with variants annotated as 'pathogenic' or 'potentially pathogenic' were excluded from further analysis.

### 4.2.2. Custom pipeline for identifying PPVs from germline WES data

The following steps were performed on exomes in the final discovery cohort following exclusion of any exomes with pathogenic variants in known CRC-predisposing genes.

#### *4.2.2.1. Transcript selection*

In order to include all variants that were potentially pathogenic, a single transcript in which an annotated variant was predicted to be most deleterious was selected for downstream analysis, as described below.

The predicted functional consequences of a variant within a transcript were ranked as shown in table 4.2. When multiple transcripts contained the same variant, the transcript with the highest ranked consequence (i.e. with the most deleterious impact) for a variant was selected for further analysis. If multiple transcripts remained with the same ranked consequence, a transcript with a RefSeq mRNA number was selected.

**Table 4-2. Ranking of predicted functional consequences of a variant**

| Predicted consequence of variant on mRNA Translation to Protein | Consequence Rank |
|---|---|
| Nonsense or frameshift | 1 |
| Affecting essential splice site | 2 |
| In-frame indel or missense | 3 |
| Stop-lost | 4 |
| Splice site, coding unknown | 5 |
| Synonymous | 6 |
| Other (not known to affect coding region) | 7 |

#### *4.2.2.2. Variant filtering strategy for identifying potentially pathogenic variants*

Consequence ranked 1-3 variants were selected for further analysis (Table 4.2), as these were more likely to be functionally deleterious. The remaining variants with a predicted consequence rank of 4 or more were omitted.

In order to prioritise PPVs in WES data that are rare, more likely to be real as opposed to artefacts, and functionally deleterious, several filters were applied. The types of filters used are summarised in Table 4.3.

**Table 4-3 Types of filters used for prioritising potentially pathogenic variants**

| Filter Criteria | Purpose |
|---|---|
| Technical filters e.g. QUAL scores, variant allele frequency | Identify alleles that are likely to be true sequence variants and not sequencing or alignment artefacts |
| Population allele frequency cut-off | Identify alleles that are rare in the population |
| In silico functional consequence scores (for missense variants) | Identify variants that are likely to be deleterious to protein function |

#### *4.2.2.3. Technical filters*

Variants had to have a QUAL score of at least 30 and the read depth at the variant locus had to be 10 or more. The variant allele frequency had to be at least 15%. While heterozygous variants are expected to have a variant allele frequency close to 50% and homozygous variants close to 100%, a permissive variant allele frequency cut-off of 15% was chosen to allow for preferential amplification of one allele over another due to sequence polymorphisms.

#### *4.2.2.4. Population filters*

##### 4.2.2.4.1. Variant filtering using population frequencies from publicly available large-scale sequencing datasets

Variants with an allele frequency of ≥0.5% in the 1000 Genomes Project, the ESP, or the Non-TCGA Subset of ExAC, were removed.

##### 4.2.2.4.2. Variant Filtering using local exome datasets

Variant counts were performed to determine the number of times each variant was seen within the sequenced exomes in the discovery set. Variants that were rare in the population but seen in 10 or more of the CRC exomes (>20%) were also excluded as they are likely to be due to sequencing or alignment errors.

A local germline exome dataset (N=237) that comprised of 72 familial breast cancer cases (Thompson et al., 2012), 57 male breast cancer cases, 41 phaeochromocytoma cases (Flynn et al., 2015), 31 mucinous ovarian tumour cases (Ryland et al., 2015), 24 serous ovarian cancer cases and 12 fibroma cases was also used for variant filtering. Variants thatwere found in at least 5% of these 237 germline exomes from exome sequencing studies performed at PMCC, Melbourne, Australia, were also excluded.

#### *4.2.2.5. Exclusion of genes with excess nonsense/frameshift variants in a local dataset*

The rationale for this filter is that genes with an excess of nonsense/frameshift variants (consequence rank 1 variants) are 'tolerant' of these variants, implying that consequence rank 1 variants in these genes are less likely to cause a disease phenotype. Using a PMCC dataset of 198 germline exomes, a count of the number of exomes with a consequence rank 1 variant in a gene, as well as a count of the total number of consequence rank 1 variants within a gene, was performed by Dr. Sally Hunter from the Campbell Lab, PMCC. This cohort of 198 germline exomes (that overlap with the local germline exome dataset from Section 4.2.2.4.2) comprised 74 male breast cancer cases, 41 phaeochromocytoma cases, 30 mucinous ovarian tumour cases, 23 low-grade serous ovarian tumour cases, 12 fibroma cases, 9 familial ovarian cancer cases, plus 9 colorectal cancer cases from the current study. Using a local exome dataset to generate this data is also useful to account for 1) polymorphisms in the local population and 2) alignment errors, as PMCC exomes are processed on the same bioinformatics variant calling pipeline. Variants in genes containing 10 or more nonsense or frameshift variants in 198 exomes (at least 1 nonsense/frameshift variant in 5% of dataset) were excluded.

#### *4.2.2.6. Prioritisation of missense variants or in-frame indels (consequence rank 3 variants) according to predicted functional impact according to CADD*

Consequence rank 3 variants that passed the preceding filters were further ranked by their CADD scaled C-score. Variants with a CADD scaled C-score ranking of >10 (corresponding to the predicted top ten percent most deleterious substitutions in the human genome), were retained for further analysis. As not all small indels were scored by CADD, in-frame indels that did not have a CADD scaled C-score were also taken forward for further analysis.

#### *4.2.2.7. Final dataset of PPVs from the discovery cohort*

All consequence rank 1, 2 and 3 variants that passed the preceding filters are henceforth referred to as PPVs. A Microsoft Excel (Microsoft Office 2010, WA, USA) pivot table with

sample and gene names on the axes was generated to visualise variants seen in a particular gene or sample/samples. The analyses described in Section 4.2.3 were performed from this pivot table. A summary of the pipeline utilising these filters is shown in Figure 4.1.

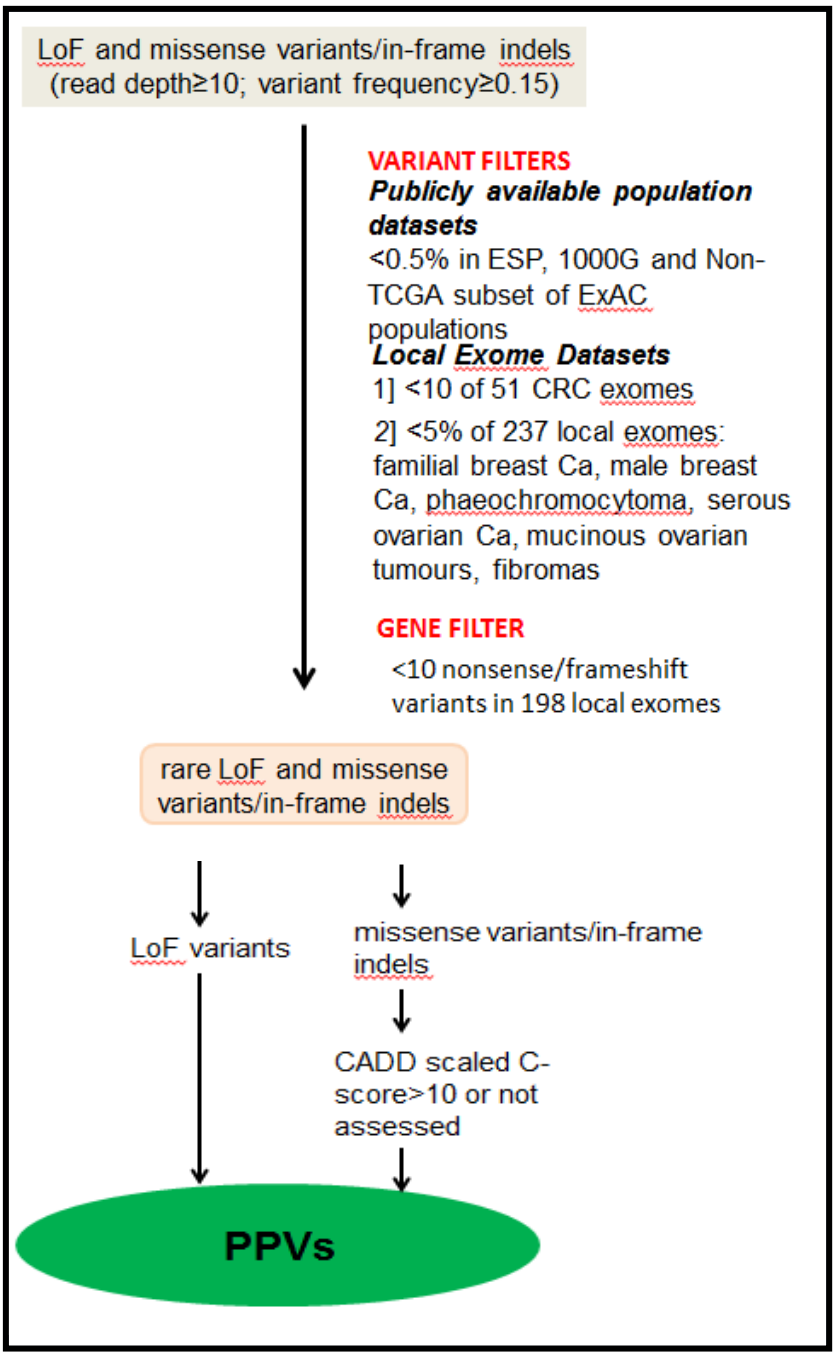



**Figure 4-1. Custom filtering pipeline for identifying potential pathogenic variants from WES data.**

**LoF, loss-of-function. ESP, Exome Sequencing Project. 1000G, 1000 Genomes Project. TCGA, The Cancer Genome Atlas. ExAC, Exome Aggregation Consortium. CADD, Combined A**

### 4.2.3. Approaches to prioritising CRC-predisposing genes harbouring potentially pathogenic variants

#### *4.2.3.1. Agnostic Approaches*

This is described in Results Sections 4.3.2.1 and 4.3.2.2.

#### *4.2.3.2. Candidate Gene Approach*

Genes within the Wnt, DNA repair, BMP and Notch signalling pathways were identified using the AmiGO web application (v.1.8) (Carbon et al., 2009) from the Gene Ontology Consortium (Ashburner et al., 2000, The Gene Ontology Consortium, 2015). The GO accession numbers corresponding to each signalling pathway and associated with the highest number of genes were selected (Table 4.4). For each selected GO accession number, Ensembl genes (Ensembl Genes 74) from the human species (Homo sapiens genes GRCh37) were retrieved on 6th February 2014. Using these gene lists, the pivot table was annotated with the relevant pathway if a pathway gene was found containing a PPV.

**Table 4-4. Gene Ontology Accession Numbers used to extract genes in a CRC pathway**

| Pathway | Gene Ontology Accession Number |
|---|---|
| Wnt | GO:0016055 |
| DNA repair | GO:0006281 |
| Bone Morphogenetic Protein | GO:0030509 |
| Notch | GO:0007219 |

#### *4.2.3.3. Gene Constraint Scores*

The known CRC-predisposing genes were annotated with their corresponding RVIS percentiles and ExAC pLI and missense z-scores in order to first evaluate their discriminatory capacity. All genes containing a PPV were then annotated with the gene constraint score with the best discriminatory capacity.

## 4.3. Results

### 4.3.1. Examination of variants in known CRC-predisposing genes in 54 exomes

The mean coverage of 11 known CRC-predisposing genes is shown in Table 4.5. With the exception of *STK11* in PC8 and PC9, two individuals with appendiceal mucinous neoplasms, the known CRC-predisposing genes had a mean coverage of at least 20X in the exomes. The low coverage of *STK11* in PC8 and PC9 is due to the use of an earlier generation (Agilent v1) exome capture reagent. All other exomes were captured using an improved Agilent v4 reagent.

**Table 4-5. Coverage of 11 known CRC-Predisposing Genes**

| Known CRC-Predisposing Genes | mean coverage in 54 exomes (range) | percentage of gene covered 1X (range) | percentage of gene covered 20X (range) |
|---|---|---|---|
| ***APC*** | 49-257 | 98-100 | 83-100 |
| ***POLE*** | 57-194 | 100 | 84-100 |
| ***POLD1*** | 42-183 | 98-100 | 63-94 |
| ***STK11*** | 2-156 | 30-100 | 4-100 |
| ***SMAD4*** | 43-241 | 100 | 82-100 |
| ***BMPR1A*** | 23-201 | 87-100 | 45-100 |
| ***MLH1*** | 46-158 | 100 | 91-100 |
| ***MSH2*** | 43-126 | 100 | 84-99 |
| ***MSH6*** | 49-257 | 100 | 86-100 |
| ***PMS2*** | 39-131 | 57-100 | 46-97 |
| ***MUTYH*** | 61-181 | 100 | 85-100 |

#### *4.3.1.1. Pathogenic LoF mutations in known CRC-predisposing genes*

PPV in the 11 known CRC-predisposing genes were inspected for pathogenicity. There were 3 pathogenic variants in known CRC-predisposing genes identified in the WES data (Table 4.6). A truncating *APC* mutation was confirmed by Sanger sequencing and reported back to the FCC. A truncating *PMS2* mutation required long-range PCR with special primers for validation. Validation of this variant was performed by the diagnostic molecular laboratory at PMCC and reported back to Professor Robyn Ward of the MCO study. A truncating *BMPR1A* mutation was reported back to the referring FCC, which performed diagnostic sequencing, confirming the presence of a pathogenic variant.

**Table 4-6. Pathogenic variants in CRC-predisposing genes in 3 individuals**

| Sample | Gene | Mutation type | Ensembl Transcript | cDNA variant | Amino acid change | Variant Frequency |
|---|---|---|---|---|---|---|
| PC15 | *PMS2* | frameshift | ENST00000265849.7 | c.736_741del6insTGTGTGTGAAG | p.Pro246Cysfs*3 | 0.43 |
| PC50 | *BMPR1A* | Stop-gained | ENST00000372037.2 | c.682C>T | p.Arg228Ter | 0.45 |
| PC51 | *APC* | frameshift | ENST00000257430.4 | c.4545delT | p.Phe1515Leufs*8 | 0.47 |

#### *4.3.1.2. Clinical context of individuals with pathogenic variants in known CRC-predisposing genes identified in WES data*

PC15 developed a sigmoid adenocarcinoma at age 31 years. According to self-reported family history, PC15's father had CRC (age unknown). IHC for *MLH1, MSH2, MSH6* and *PMS2* did not show any absence of staining in the tumour. WES of germline DNA from PC15 identified a complex indel in *PMS2*, ENST00000265849.7:c.736_741del6insTGTGTGTGAAG, p.Pro246Cysfs*3. This complex indel has been reported previously, in 12 of 61 patients (20%) with a *PMS2* mutation and isolated PMS2 loss on MMR IHC of their cancer (Clendenning et al., 2008). It is thought to be a founder mutation. It is unclear why IHC for PMS2 was positive in PC15. The specific antibody used for IHC in this case was not known, hence it is unclear if the N-terminus or C-terminus was targeted. Tumour from PC15 was not available for LOH analysis. This PMS2 mutation is usually associated with a negative result on IHC for PMS2 in the literature, although this is probably because *PMS2* mutation cases are frequently ascertained by IHC. There are examples of *PMS2* exon deletions with retained PMS2 expression on IHC in the literature (van der Klift et al., 2016).

PC50 had a history of per rectum bleeding at the age of 14 years and colonic polyposis (polyp type unknown, thought to be mainly adenomas). The patient underwent a total colectomy overseas, prior to migrating to Australia. The histopathology report from the subtotal colectomy was not available. The patient developed a rectal cancer at age 48 years. Two of the patient's 11 siblings developed colonic polyposis and underwent colectomies overseas in their home country. One of these affected siblings had a son who had colonic polyposis and CRC at age 22 and unfortunately died from the disease. Three of the patient's siblings, including one with colonic polyposis, died in childhood. The patient's parents had no history of CRC, with the patient's father having had a normal colonoscopy and dying in his 70s. The patient's mother is alive in her 80s and has not had a colonoscopy. Testing of *APC* was performed for the patient, initially with Protein Truncation Test (PTT) of exon 15 and denaturing high performance liquid chromatography of exons 1-14. No mutations were found. Full gene sequencing and MLPA of *APC* and *MUTYH* were performed between 2008 and 2009, and was uninformative. Whole exome sequencing of germline DNA identified a stop-gained mutation in *BMPR1A*, ENST00000372037.2: c.682C>T, p.Arg228Ter.

PC51 had a clinical picture and family history consistent with FAP. WES of blood DNA subsequently identified a pathogenic frameshift mutation (ENST00000257430.4 (*APC*):

c.4545delT). It is unclear why this mutation was missed on PTT. The PTT report specified that 5 fragments were analysed, and no abnormal sized protein was seen. Reference was not given to the specific PTT protocol used. The most common reasons for a negative PTT are absence of the mutated allele due to failure of amplification, possibly due to an altered primer binding site, or mutations at the very start or end of the protein (Den Dunnen and Van Ommen, 1999). APC is 2843 amino acids long, hence this mutation is not at the extreme ends of the protein. This mutation is in Exon 15, the largest exon. This exon is amplified from DNA, not RNA, hence RNA degradation due to nonsense-mediated decay is not a factor here. Overlap of the 5 fragments for PTT should ensure that mutations near the ends of a fragment are not missed (Powell et al., 1993).

#### *4.3.1.3. Rare variants of uncertain significance in known CRC-predisposing genes*

Heterozygous missense variants and an essential splice site variant of uncertain significance identified in known CRC-predisposing genes in WES data are summarised in Table 4.7. These variants were found in *APC, MLH1, MSH6, POLE* and *POLD1*. The essential splice site variant in *MSH6* (ENST00000234420.4:c.4001+2_4001+5delTAAC) did not alter the essential splice site region of *MSH6* due to the presence of TAAC repeats. As MSH6 was present on IHC in the tumour from PC33, this variant is unlikely to be pathogenic.

The associated in silico predictors of pathogenicity, SIFT, PolyPhen, Condel and CADD are shown for the variants of uncertain significance. None of these variants were seen in ClinVar and therefore these cases were retained for analysis in the discovery cohort.

**Table 4-7. Variants of uncertain significance in known CRC-predisposing genes**

| Sample | Variant Consequence | Gene | Ensembl Transcript Consequence | PolyPhen | SIFT | Condel | CADD Scaled Score | ClinVar annotation |
|---|---|---|---|---|---|---|---|---|
| **PC4** | missense | *APC* | ENSP00000457016.1:p.Ala1879Ser | benign (0.017) | tolerated (0.07) | neutral (0.320) | 14.69 | . |
| **PC21** | missense | *APC* | ENSP00000457016.1:p.Gln203Glu | possibly_damaging (0.807) | deleterious (0.01) | deleterious (0.710) | 29.9 | . |
| **PC25** | missense | *APC* | ENSP00000457016.1:p.Leu1129Ser | benign(0.357) | tolerated(0.1) | neutral(0.399) | 12.69 | . |
| **PC39** | missense | *APC* | ENSP00000457016.1:p.Arg99Trp | probably_damaging (0.949) | deleterious (0.02) | deleterious (0.776) | 19.71 | . |
| **PC49** | missense | *APC* | ENSP000004570160.1:p.Arg2505Gln | possibly_damaging (0.868) | tolerated (0.26) | neutral (0.371) | 35 | . |
| **PC29** | missense | *MLH1* | ENSP00000231790.2:p.Gly22Ala | probably_damaging (0.988) | deleterious (0.01) | deleterious (0.843) | 34 | . |
| **PC33** | essential splice site | *MSH6* | ENST00000234420.4:c.4001+2_4001+5delTAAC | . | . | . | 16.29 | . |
| **PC39** | missense | *POLE* | ENSP00000322570.5:p.Ala724Val | benign (0.269) | tolerated (0.14) | neutral (0.278) | 21 | . |
| **PC43** | missense | *POLE* | ENSP00000322570.5:p.Gly1352Ala | benign (0.048) | tolerated (0.09) | neutral (0.299) | 15.74 | . |
| **PC8, PC9** | missense | *POLE* | ENSP00000322570.5:p.Val1887Met | benign (0.202) | tolerated (0.29) | neutral (0.049) | 10.07 | . |
| **PC48** | missense | *POLE* | ENSP00000322570.5:p.Ala1420Val | benign (0.006) | tolerated (0.26) | neutral (0.034) | 13.99 | . |
| **PC11** | missense | *POLD1* | ENSP00000406046.1:p.Glu673Lys | possibly_damaging (0.598) | deleterious (0) | deleterious (0.651) | 33 | . |

Three samples (PC15, PC50 and PC51) were excluded from further analysis following the finding of pathogenic variants in known CRC-predisposing genes. The final discovery cohort comprised 51 exomes.

### 4.3.2. Performance summaries for 51 exomes

The sequencing performance summaries for the 51 exomes are summarised in Table 4.8. Between 62.4 million and 135.8 million reads were generated per exome. On average 99% of reads were mapped to the reference sequence. All but one exome (PC1) had less than 20% of mapped reads as sequence duplicates, which were removed after alignment. The percentage of reads on target ranged from 39% to 79%, with an average of 68%. The percentage of bases covered at least 10 times was more than 95% for all exomes, and the percentage of bases covered at least 20 times was more than 85% for 50 exomes, and 84.18% for PC47. The mean coverage of target bases per exome ranged from 49 times to 136 times, with a mean of 83 times across the 51 exomes.

**Table 4-8. Sequencing performance summaries of 51 exomes in the discovery cohort**

| Sample | Exome Enrichment Method | Sequence reads | Read length (bp) | Total reads | % Reads mapped | % Mapped reads duplicates | % Reads OnTarget | % Target bases >=10-fold Coverage | % Target bases >=20-fold Coverage | Mean coverage for target bases |
|---|---|---|---|---|---|---|---|---|---|---|
| PC1 | Agilent SureSelect Human All Exon v4 | Paired-end | 100 | 126651040 | 89.44 | 27.23 | 66.94 | 98.21 | 93.79 | 93.72 |
| PC2 | Agilent SureSelect Human All Exon v4 | Paired-end | 100 | 107973154 | 97.66 | 14.14 | 71.81 | 98.01 | 93.33 | 109.26 |
| PC3 | Agilent SureSelect Human All Exon v4 | Paired-end | 100 | 135889460 | 97.23 | 13.36 | 71.1 | 98.46 | 94.86 | 136.87 |
| PC4 | Agilent SureSelect Human All Exon v4 | Paired-end | 100 | 122936952 | 96.4 | 13.95 | 72.76 | 97.71 | 92.68 | 117.38 |
| PC5 | Agilent SureSelect Human All Exon v4 | Paired-end | 100 | 81405842 | 98.33 | 8.37 | 69.39 | 96.26 | 88.04 | 79.28 |
| PC6 | Agilent SureSelect Human All Exon v4 | Paired-end | 100 | 84656078 | 98.21 | 6.08 | 70.67 | 96.48 | 89.05 | 86 |
| PC7 | Agilent SureSelect Human All Exon v4 | Paired-end | 100 | 111736018 | 93.51 | 16.03 | 75.84 | 97.9 | 93.28 | 114.64 |
| PC8 | Agilent SureSelect Human All Exon v1 | Paired-end | 100 | 65504122 | 99.98 | 4.81 | 77.6 | 95.42 | 91.04 | 130.69 |
| PC9 | Agilent SureSelect Human All Exon v1 | Paired-end | 100 | 70511080 | 99.98 | 12.4 | 75.41 | 95.71 | 90.34 | 126.95 |
| PC10 | Agilent SureSelect Human All Exon v4 | Paired-end | 100 | 68569336 | 99.96 | 6 | 79.12 | 98.05 | 94.84 | 76.64 |
| PC11 | Agilent SureSelect Human All Exon v4 | Paired-end | 100 | 68517184 | 99.95 | 6.2 | 71.47 | 97.41 | 93.55 | 69.25 |
| PC12 | Agilent SureSelect Human All Exon v4 | Paired-end | 100 | 68723634 | 99.96 | 6.64 | 73.74 | 97.81 | 93.99 | 71.44 |
| PC13 | Agilent SureSelect Human All Exon v4 | Paired-end | 100 | 64981092 | 99.95 | 3.99 | 54.87 | 97.16 | 88.44 | 51.92 |
| PC14 | Agilent SureSelect Human All Exon v4 | Paired-end | 100 | 63310956 | 99.97 | 4.6 | 75.37 | 98.17 | 93.34 | 69.37 |
| PC16 | Agilent SureSelect Human All Exon v4 | Paired-end | 100 | 64716704 | 99.95 | 6.51 | 53.95 | 96.92 | 87.12 | 49.4 |
| PC17 | Agilent SureSelect Human All Exon v4 | Paired-end | 100 | 62499640 | 99.97 | 7.32 | 73.4 | 98.13 | 92.72 | 64.39 |
| PC18 | Agilent SureSelect Human All Exon v4 | Paired-end | 100 | 68377119 | 99.94 | 5.86 | 64.93 | 97.57 | 92.45 | 62.87 |

**Table 4-8 (continued). Sequencing performance summaries of 51 exomes in the discovery cohort**

| Sample | Exome Enrichment Method | Sequence reads | Read length (bp) | Total reads | % Reads mapped | % Mapped reads duplicates | % Reads OnTarget | % Target bases >=10-fold Coverage | % Target bases >=20-fold Coverage | Mean coverage for target bases |
|---|---|---|---|---|---|---|---|---|---|---|
| **PC19** | Agilent SureSelect Human All Exon v4 | Paired-end | 100 | 64786809 | 99.95 | 4.25 | 53.63 | 97.07 | 87.5 | 50.31 |
| **PC20** | Agilent SureSelect Human All Exon v4 | Paired-end | 100 | 68090418 | 99.96 | 6.9 | 77.97 | 98.16 | 94.51 | 74.7 |
| **PC21** | Agilent SureSelect Human All Exon v4 | Paired-end | 100 | 78733319 | 99.87 | 4.35 | 67.78 | 98.52 | 94.84 | 76.35 |
| **PC22** | Agilent SureSelect Human All Exon v4 | Paired-end | 100 | 74222285 | 99.83 | 3.15 | 56.98 | 98.01 | 92.48 | 61.34 |
| **PC23** | Agilent SureSelect Human All Exon v4 | Paired-end | 100 | 87842858 | 99.81 | 7.04 | 52.38 | 97.67 | 92.98 | 63.54 |
| **PC24** | Agilent SureSelect Human All Exon v4 | Paired-end | 100 | 93638124 | 99.89 | 5.38 | 67.38 | 99.02 | 97.01 | 90.04 |
| **PC25** | Agilent SureSelect Human All Exon v4 | Paired-end | 100 | 67616062 | 99.87 | 4.72 | 67.66 | 98.23 | 94.07 | 65.39 |
| **PC26** | Agilent SureSelect Human All Exon v4 | Paired-end | 100 | 68758181 | 99.91 | 4.14 | 66.64 | 98.35 | 94.22 | 65.81 |
| **PC27** | Agilent SureSelect Human All Exon v4 | Paired-end | 100 | 87335153 | 99.92 | 5.52 | 68.67 | 98.73 | 96.37 | 84.55 |
| **PC28** | Agilent SureSelect Human All Exon v4 | Paired-end | 100 | 63177103 | 99.9 | 3.67 | 69.63 | 98.15 | 93.58 | 63.78 |
| **PC29** | Agilent SureSelect Human All Exon v4 | Paired-end | 100 | 78264971 | 99.9 | 4.62 | 72.52 | 98.58 | 95.8 | 81.62 |
| **PC30** | Agilent SureSelect Human All Exon v4 | Paired-end | 100 | 81408188 | 99.78 | 8.38 | 54.21 | 97.59 | 91.94 | 60.54 |
| **PC31** | Agilent SureSelect Human All Exon v4 | Paired-end | 100 | 94254595 | 99.88 | 8.93 | 53.95 | 98.14 | 94.3 | 69.96 |
| **PC32** | Agilent SureSelect Human All Exon v4 | Paired-end | 100 | 88195930 | 99.85 | 7.4 | 49.4 | 97.67 | 92.35 | 60.87 |
| **PC33** | Agilent SureSelect Human All Exon v4 | Paired-end | 100 | 83881990 | 99.94 | 9.81 | 73.37 | 98.84 | 96.09 | 91.42 |
| **PC34** | Agilent SureSelect Human All Exon v4 | Paired-end | 100 | 97963726 | 99.94 | 11.05 | 72.92 | 99.07 | 97.14 | 104.91 |
| **PC35** | Agilent SureSelect Human All Exon v4 | Paired-end | 100 | 102001749 | 99.7 | 9.35 | 74.12 | 98.87 | 97.19 | 113.91 |
| **PC36** | Agilent SureSelect Human All Exon v4 | Paired-end | 100 | 88916340 | 99.95 | 10.2 | 71.66 | 98.88 | 96.39 | 94.49 |
| **PC37** | Agilent SureSelect Human All Exon v4 | Paired-end | 100 | 95999995 | 99.95 | 10.13 | 71.7 | 99.01 | 96.84 | 101.99 |
| **PC38** | Agilent SureSelect Human All Exon v4 | Paired-end | 100 | 97792204 | 99.95 | 11.61 | 73.57 | 99.08 | 97.08 | 105.22 |
| **PC39** | Agilent SureSelect Human All Exon v4 | Paired-end | 100 | 99994030 | 99.95 | 5.45 | 68.2 | 98.87 | 96.99 | 106.12 |
| **PC40** | Agilent SureSelect Human All Exon v4 | Paired-end | 100 | 95233031 | 99.89 | 5.74 | 72.41 | 98.64 | 96.27 | 106.75 |
| **PC41** | Agilent SureSelect Human All Exon v4 | Paired-end | 100 | 88351876 | 99.94 | 5.71 | 67.33 | 98.64 | 96.04 | 92.13 |

**Table 4-8 (continued). Sequencing performance summaries of 51 exomes in the discovery cohort**

| Sample | Exome Enrichment Method | equence reads | Read length (bp) | Total reads | % Reads mapped | % Mapped reads duplicates | % Reads OnTarget | % Target bases >=10-fold Coverage | % Target bases >=20-fold Coverage | Mean coverage for target bases |
|---|---|---|---|---|---|---|---|---|---|---|
| **PC42** | Agilent SureSelect Human All Exon v4 | Paired-end | 100 | 83140403 | 99.94 | 10.33 | 73.18 | 98.71 | 96.03 | 90.03 |
| **PC43** | Agilent SureSelect Human All Exon v4 | Paired-end | 100 | 69682847 | 99.88 | 10.01 | 73.11 | 98.35 | 94.19 | 75.24 |
| PC44 | Agilent SureSelect Human All Exon v4 | Paired-end | 100 | 92736224 | 90.84 | 11.27 | 75.93 | 96.45 | 89.6 | 98.13 |
| PC45 | Agilent SureSelect Human All Exon v4 | Paired-end | 100 | 89910941 | 99.66 | 2.93 | 39.65 | 97.23 | 89.44 | 51.44 |
| PC46 | Agilent SureSelect Human All Exon v4 | Paired-end | 100 | 68994042 | 99.97 | 6.33 | 79.1 | 98.01 | 94.71 | 77.01 |
| PC47 | Agilent SureSelect Human All Exon v4 | Paired-end | 100 | 65682485 | 99.96 | 7.67 | 76.85 | 95.9 | 84.18 | 71.09 |
| PC48 | Agilent SureSelect Human All Exon v4 | Paired-end | 100 | 68502016 | 99.96 | 6.84 | 79.14 | 98.05 | 94.73 | 76.26 |
| **PC49** | Agilent SureSelect Human All Exon v4 | Paired-end | 100 | 97355157 | 99.79 | 10.04 | 71.27 | 98.98 | 96.92 | 102.64 |
| PC52 | Agilent SureSelect Human All Exon v4 | Paired-end | 100 | 65537454 | 99.97 | 4.8 | 75.65 | 96.97 | 88.4 | 72.01 |
| **PC53** | Agilent SureSelect Human All Exon v4 | Paired-end | 100 | 85282696 | 99.77 | 8.28 | 51.76 | 96.78 | 89.07 | 60.3 |
| **PC54** | Agilent SureSelect Human All Exon v4 | Paired-end | 100 | 88396817 | 99.88 | 18.98 | 73.67 | 98.34 | 95.75 | 86.29 |

### 4.3.3. Distribution of potentially pathogenic variants per exome

The median number of LoF variants per exome was 19 (range 11-92) (Table 4.9). The outlier sample (PC47) with 92 LoF variants consisted mostly of indels (88%). Approximately two-thirds of the indels had low variant allele frequencies (68% ≤0.2), and were not reported in the dbSNP database (Sherry et al., 2001). These additional variants are likely due to errors in sequencing and they were kept in for the intra-family analysis, but not included in the inter-family analyses.

**Table 4-9. Range of rare LoF variant counts per exome**

| Number of LoF variants | Number of exomes |
|---|---|
| 11-20 | 30 |
| 21-30 | 18 |
| 31-45 | 2 |
| >45 | 1 |

The number of potentially pathogenic missense variants per exome is shown in Table 4.10. There were 10609 missense variants and 421 in-frame indels. The median number of missense variants/in-frame indels per exome was 202 (range 137-588). Three exomes had more than 300 missense variants/in-frame indels (PC28, PC1 and PC27 had 313, 321 and 588 variants respectively). In PC27, the sample with 588 missense variants/in-frame indels, 202 of these variants have an allele frequency of 1-2% in the African population within the 1000 Genomes Project, indicating that the excess number of variants in this individual is due to population stratification. In contrast, in the other 50 exomes, the maximum number of missense variants/in-frame indels with an allele frequency of 1-2% in the African population within the 1000 Genomes Project was 6.

**Table 4-10. Range of rare missense variant/in-frame indel counts per exome**

| Number of missense variants/in-frame indels | Number of exomes |
|---|---|
| 137-200 | 24 |
| 201-300 | 24 |
| >300 | 3 |

### 4.3.4. Number of variants at each data filtering step

The 51 exomes comprising the discovery cohort contained a total of 58 977 LoF and missense variants/in-frame indels at genomic locations covered by at least 10 reads, with a variant allele frequency of at least 15%. Following population filters with the 1000 Genomes Project, ESP and Non-TCGA Subset of ExAC as described above, 26 971 rare variants remained. Further filtering using the 51 CRC exomes themselves and the PMCC multi-project exomes (N=237) reduced the number of LoF and missense variants to 17 950, in 8695 genes. 1182 variants were excluded as they were in genes that were deemed from a local exome dataset of 198 germline exomes to have an excess of nonsense or frameshift variants and likely to be artefacts. After this filtering, 1090 LoF and 15 678 missense variants/in-frame indels remained. Of the 15 678 missense variants/in-frame indels, 11 030 had a CADD scaled score of more than 10 or were not able to be ranked by CADD and were kept for further analysis. All 1090 LoF variants were retained for further analysis.

### 4.3.5. Prioritisation of candidate CRC-predisposing genes containing PPVs

Having identified PPVs through a filtering pipeline described in Section 4.2.2, evidence that these variants may predispose to CRC was sought. The following 4 analyses were performed:

1) Inter-family (Section 4.3.5.1.1)

2) Intra-family (Section 4.3.5.1.2)

3) Genes in CRC pathways (4.3.5.2)

4) Gene constraint scores (4.3.5.3)

Analysis was performed at the gene level for analysis methods 1, 3 and 4, and at a variant level for analysis method 2.

#### *4.3.5.1. Agnostic Approach*

##### 4.3.5.1.1. Inter-family analysis

Ninety genes harboured LoF variants in individuals from at least 2 families. Two hundred and fifty-four genes harboured a LoF variant in one family and one or more missense variants/in-frame indels in a different family. Two thousand one hundrend and twenty-eight genes harboured a missense variant in at least 2 families. Subsequent analysis was restricted to LoF variants, as these variants were most likely to be deleterious. Stringent quality thresholds were applied to selecting LoF variants of interest, in order to prioritise variants that were most likely to be real and de-prioritise sequencing or alignment artefacts. The base calling Phred quality score (QUAL), variant frequency and sequence read direction of each variant was examined. Variants that fulfilled two of three of the following criteria were omitted: 1) QUAL score<100, 2) unidirectional rather than bidirectional supporting reads, or 3) variant frequency of ≤0.3. These empirical cut-offs were chosen to prioritise only the variants with very high supporting quality metrics. The population allele frequency cut-off was decreased from 0.5% to <0.1% in the Non-TCGA Subset of ExAC in order to prioritise rarer variants. Genes encoding olfactory receptors were omitted, as these genes are known to tolerate a high number of LoF variants (Lek et al., 2016c). All remaining LoF variants were visualised using the Integrated Genome Viewer (Robinson et al., 2011). Multi-allelic variants, variants in repetitive regions or regions with multiple low quality reads were excluded.

Following these filtering steps, 12 genes with a LoF variant in different families were found in the discovery dataset. Each gene had a LoF variant in 2 unrelated individuals (Table 4.11). For 5 genes, the LoF variant was identical in the 2 individuals.

**Table 4-11. Genes with LoF variants in 2 unrelated individuals in the discovery cohort**

| Gene | Gene description | Ensembl Gene | Genomic position | Ensembl transcript | cDNA position | Amino acid change | ExAC-TCGA AF | Same variant? | Sample |
|---|---|---|---|---|---|---|---|---|---|
| ***SERPINB7*** | serpin peptidase inhibitor, clade B (ovalbumin), member 7 | ENSG00000166396 | 18:61468141_61468142del GT | ENST00000398019.2 | c.639_640delGT | p.Lys213AsnfsTer7 | . | No | PC20 |
| ***SERPINB7*** | serpin peptidase inhibitor, clade B (ovalbumin), member 7 | ENSG00000166396 | 18:61471588A>T | ENST00000398019.2 | c.862A>T | p.Lys288Ter | . | No | PC29 |
| ***GALC*** | galactosylceramidase | ENSG00000054983 | 14:88459424_88459425insCGCGG | ENST00000261304.2 | c.84_85insCCGCG | p.Val29ProfsTer45 | . | No | PC1 |
| ***GALC*** | galactosylceramidase | ENSG00000054983 | 14:88459803_88459804insG | ENST00000393569.2 | c.40_41insC | p.Leu14ProfsTer14 | . | No | PC28 |
| ***GALNT8*** | UDP-N-acetyl-alpha-D-galactosamine:polypeptide N-acetylgalactosaminyltransferase 8 (GalNAc-T8) | ENSG00000130035 | 12:4870307C>T | ENST00000252318.2 | c.1357C>T | p.Gln453Ter | 1.22E-04 | No | PC11 |
| ***GALNT8*** | UDP-N-acetyl-alpha-D-galactosamine:polypeptide N-acetylgalactosaminyltransferase 8 (GalNAc-T8) | ENSG00000130035 | 12:4872418G>A | ENST00000252318.2 | c.1360-1G>A | | . | No | PC3 |
| ***GPR144*** | G protein-coupled receptor 144 | ENSG00000180264 | 9:127215935G>A | ENST00000334810.1 | c.959G>A | p.Trp320Ter | 2.39E-04 | No | PC33 |
| ***GPR144*** | G protein-coupled receptor 144 | ENSG00000180264 | 9:127239323_127239324delCT | ENST00000334810.1 | c.2836_2837del CT | p.Phe947SerfsTer4 | . | No | PC44 |
| ***PDE5A*** | phosphodiesterase 5A, cGMP-specific | ENSG00000138735 | 4:120423760_120423763 delTCTC | ENST00000264805.5 | c.2253_2256del GAGA | p.Arg752LysfsTer5 | . | No | PC22 |
| ***PDE5A*** | phosphodiesterase 5A, cGMP-specific | ENSG00000138735 | 4:120548312T>A | ENST00000264805.5 | c.19A>T | p.Lys7Ter | 1.51E-04 | No | PC17 |
| ***MS4A15*** | membrane-spanning 4-domains, subfamily A, member 15 | ENSG00000166961 | 11:60538775delG | ENST00000337911.4 | c.82delG | p.Gly28AspfsTer16 | 1.88E-05 | No | PC33 |
| ***MS4A15*** | membrane-spanning 4-domains, subfamily A, member 15 | ENSG00000166961 | 11:60541424G>C | ENST00000337911.4 | c.333+1G>C | . | 1.88E-05 | No | PC17 |
| ***ZNF788*** | Zinc finger protein 788 | ENSG00000188474 | 19:12223126delG | ENST00000339302.4 | c.765delG | p.Glu256AsnfsTer13 | 2.85E-04 | No | PC17 |
| ***ZNF788*** | Zinc finger protein 788 | ENSG00000188474 | 19:12223499_12223500delAT | ENST00000339302.4 | c.1137_1138del AT | p.Cys380Ter | 9.47E-05 | No | PC23 |
| ***RIPPLY3*** | ripply3 homolog (zebrafish) | ENSG00000183145 | 21:38390226C>T | ENST00000329553.2 | c.292C>T | p.Gln98Ter | 3.39E-04 | Yes | PC12, PC29 |
| ***ASCL3*** | achaete-scute complex homolog 3 (Drosophila) | ENSG00000176009 | 11:8959327G>A | ENST00000325884.1 | c.382C>T | p.Arg128Ter | 7.16E-04 | Yes | PC36, PC17 |
| ***TEX14*** | testis expressed 14 | ENSG00000121101 | 17:56690784G>A | ENST00000349033.5 | c.1003C>T | p.Arg335Ter | 8.38E-04 | Yes | PC49, PC20 |
| ***SH2D4A*** | SH2 domain containing 4A | ENSG00000104611 | 8:19218826G>A | ENST00000265807.3 | c.706+1G>A | . | 9.60E-04 | Yes | PC40, PC10 |
| ***TNFRSF10B*** | tumor necrosis factor receptor superfamily, member 10b | ENSG00000120889 | 8:22885843C>T | ENST00000276431.4 | c.748+1G>A | . | 3.01E-04 | Yes | PC20, PC30 |

ExAC-TCGA AF, variant allele frequencies from the Non-TCGA Subset of ExAC.

#### 4.3.5.1.2. Intra-family analyses

Five families were analysed in further detail, as 2 affected individuals from each of these 5 families consented to WES (Figures 4.2 to 4.5). These families consisted of 4 pairs of FDR, and one pair of SDR. The probands are indicated with a black triangle on the pedigrees, and the age of onset of CRC is also shown. The pedigrees and analysis for 4 families is described below. The pedigree and analysis for the family with AMTs is described in the accompanying manuscript 'Candidate Genes Predisposing to Familial Appendiceal Mucinous Tumours Presenting With Pseudomyxoma Peritonei' (4.3.5.1.2.4). Familial AMT was analysed using more stringent population filters as the incidence of AMT is much lower than CRC in general. Genome-wide copy number analysis of tumour tissue from the affected father and daughter was also performed to identify areas of LOH within the tumour, as described in the manuscript.

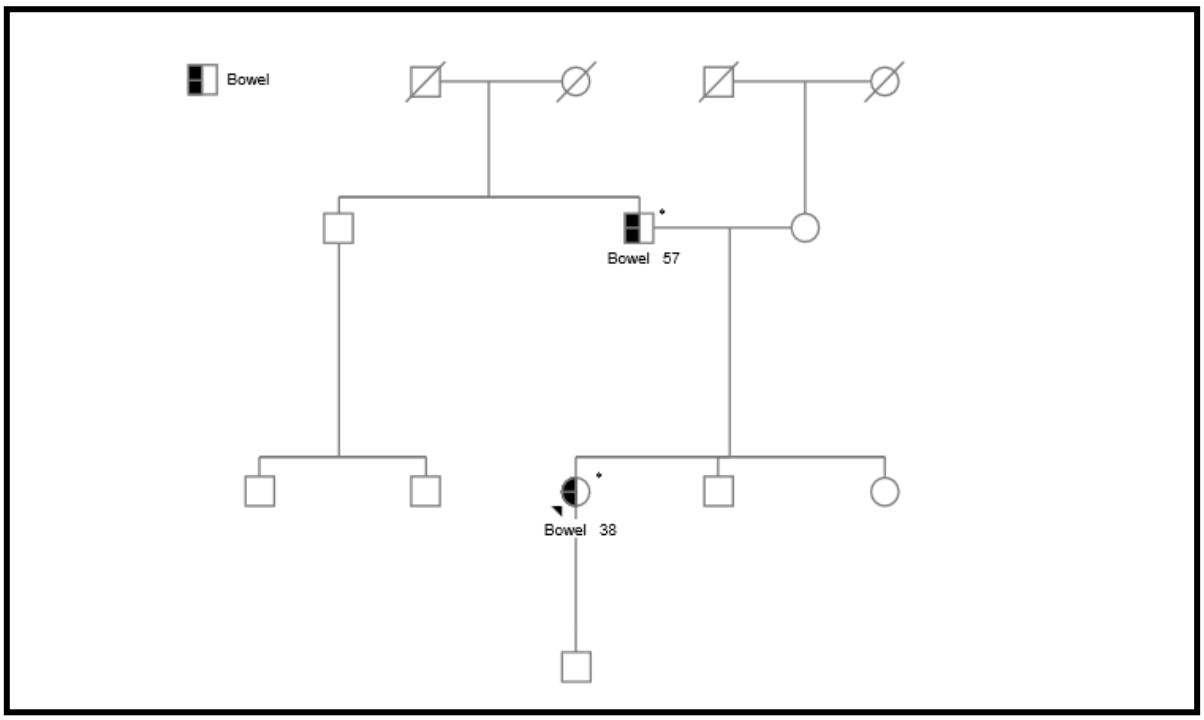



**Figure 4-2. Pedigree of father (PC3) and daughter (PC2) with rectal adenocarcinoma**

The daughter (proband) was diagnosed with a rectal adenocarcinoma at the age of 38 years that was mismatch repair proficient and microsatellite stable. The father was diagnosed with a rectal adenocarcinoma at the age of 57 years.

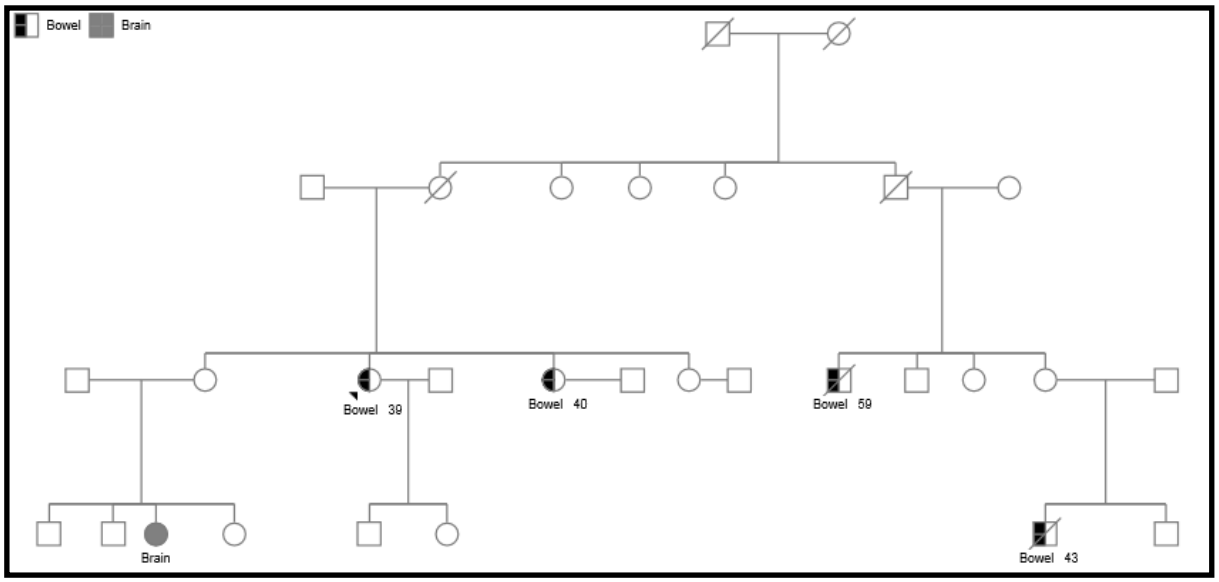



**Figure 4-3. Pedigree of sisters (PC6 and PC7) with young-onset CRC**

The proband (PC6) was diagnosed with CRC at age 39 years. Her sister (PC7) was diagnosed with a CRC at the age of 40 years. While PC6's CRC was mismatch repair proficient, PC7's CRC was deficient of MSH2 and MSH6 on immunohistochemistry. Germline sequencing and MLPA of the *MSH2* and *MSH6* genes in PC8 did not reveal any pathogenic variants.

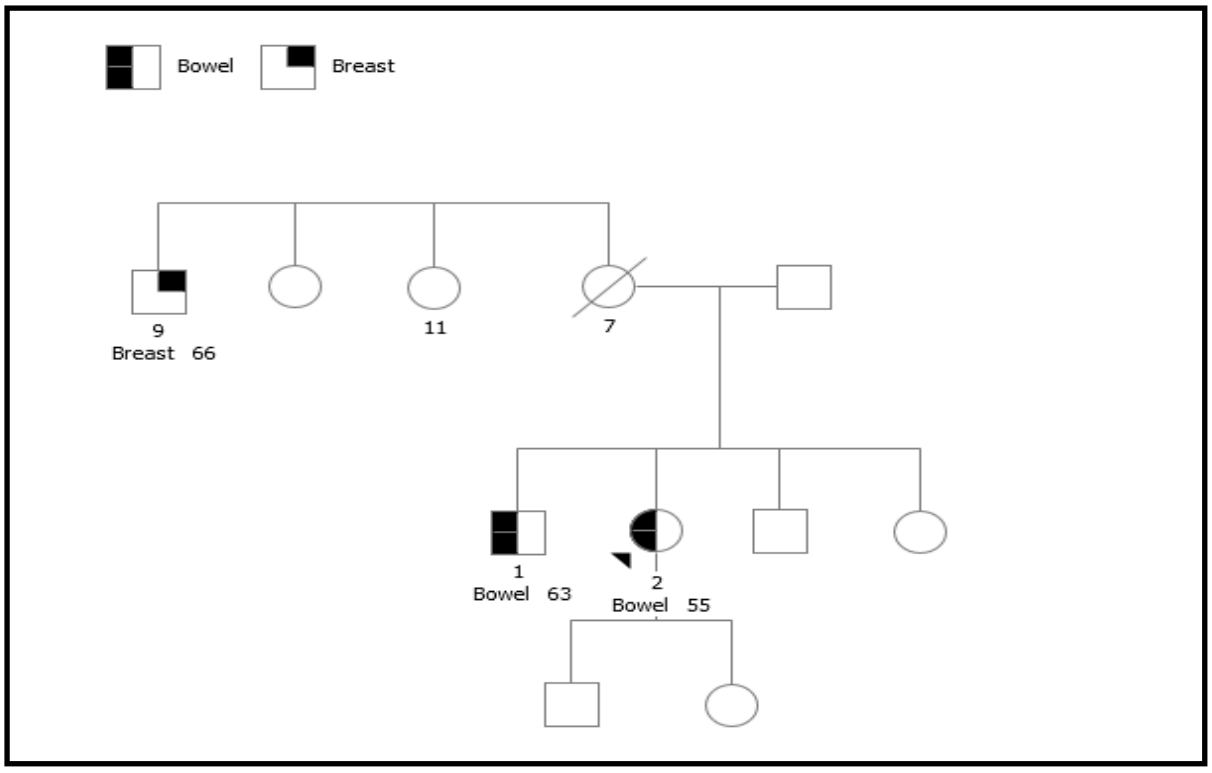



**Figure 4-4. Pedigree of brother (PC46) and sister (PC47) with colonic polyposis and CRC**

The proband (PC47) had a positive screening faecal occult blood test at the age of 55 years. On colonoscopy, she was found to have a rectal adenocarcinoma and approximately 80 colonic

polyps, that consisted of tubular adenomas, sessile serrated polyps and hyperplastic polyps. Her asymptomatic brother (PC46) underwent colonoscopy and was found to have a rectosigmoid adenocarcinoma and multiple colonic polyps. On subtotal colectomy PC46 had 25 sessile polyps throughout the colon, consisting of hyperplastic polyps, tubular adenomas and tubulovillous adenomas with low grade dysplasia, in addition to the rectosigmoid adenocarcinoma.

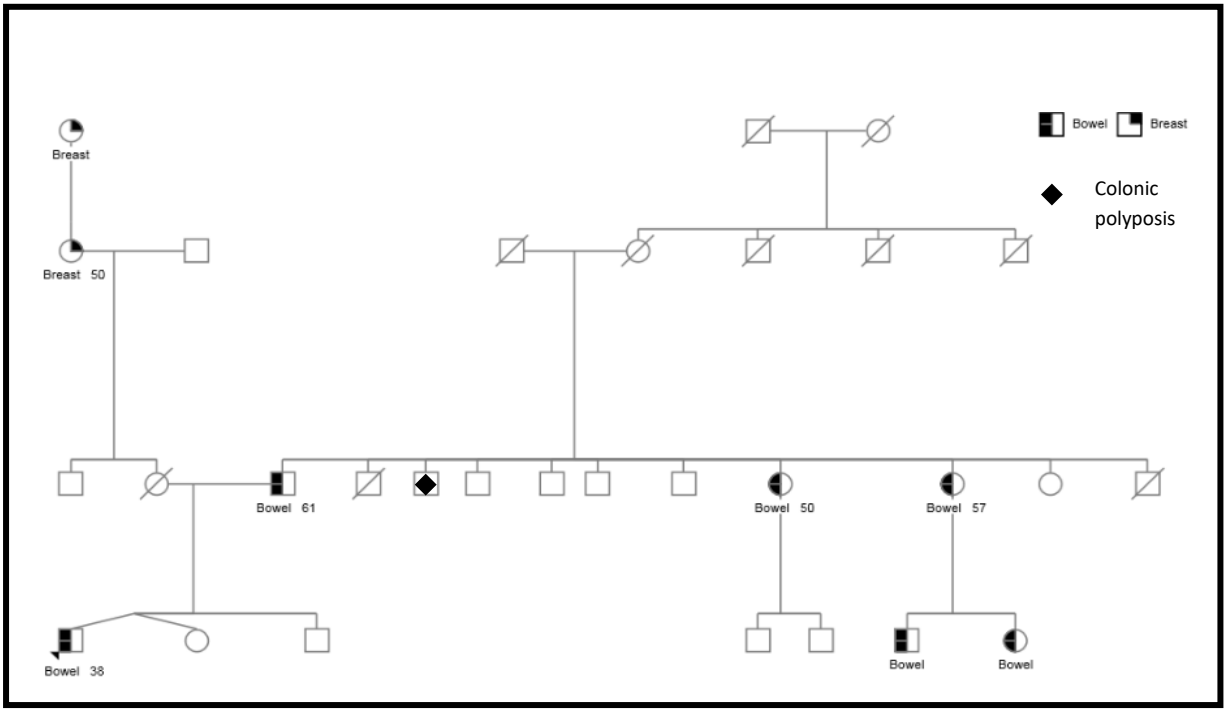



**Figure 4-5. Pedigree of paternal uncle (PC54) with colonic adenomatous polyposis and nephew (PC41) with young-onset CRC. PC54 developed colonic polyposis without CRC.**

This family has familial colorectal cancer type X, as it fulfils Amsterdam I criteria and the tumour of the proband (PC41) was proficient in the mismatch repair proteins MLH1, MSH2, MSH6 and PMS2 tested by immunohistochemistry. PC41's tumour was also microsatellite stable at 5 mononucleotide microsatellite markers. PC41 was diagnosed with a caecal adenocarcinoma at the age of 38 years. PC54 had 41 tubular or tubulovillous adenomas removed from his colon between the ages of 56 and 73 years. PC54 was chosen for WES rather than the affected father of PC41 to reduce the number of shared variants. The affected paternal aunts and cousins of PC41 were invited to participate in the study but did not respond to the invitation.

#### *4.3.5.1.2.1. Rare LoF and missense/in-frame indel variants found in both affected family members within a family*

Using the rare variants identified in Section 4.2.1, the intra-family analysis for 4 of the above families is described. Identical rare LoF and missense variants that were identified in both affected individuals from a family were prioritised. There were between 4 and 15 rare LoF variants shared between the 2 affected individuals within each family. As was done in the

inter-family analysis, further filters were applied to prioritise the variants most likely to predispose to CRC (Figure 4-6).

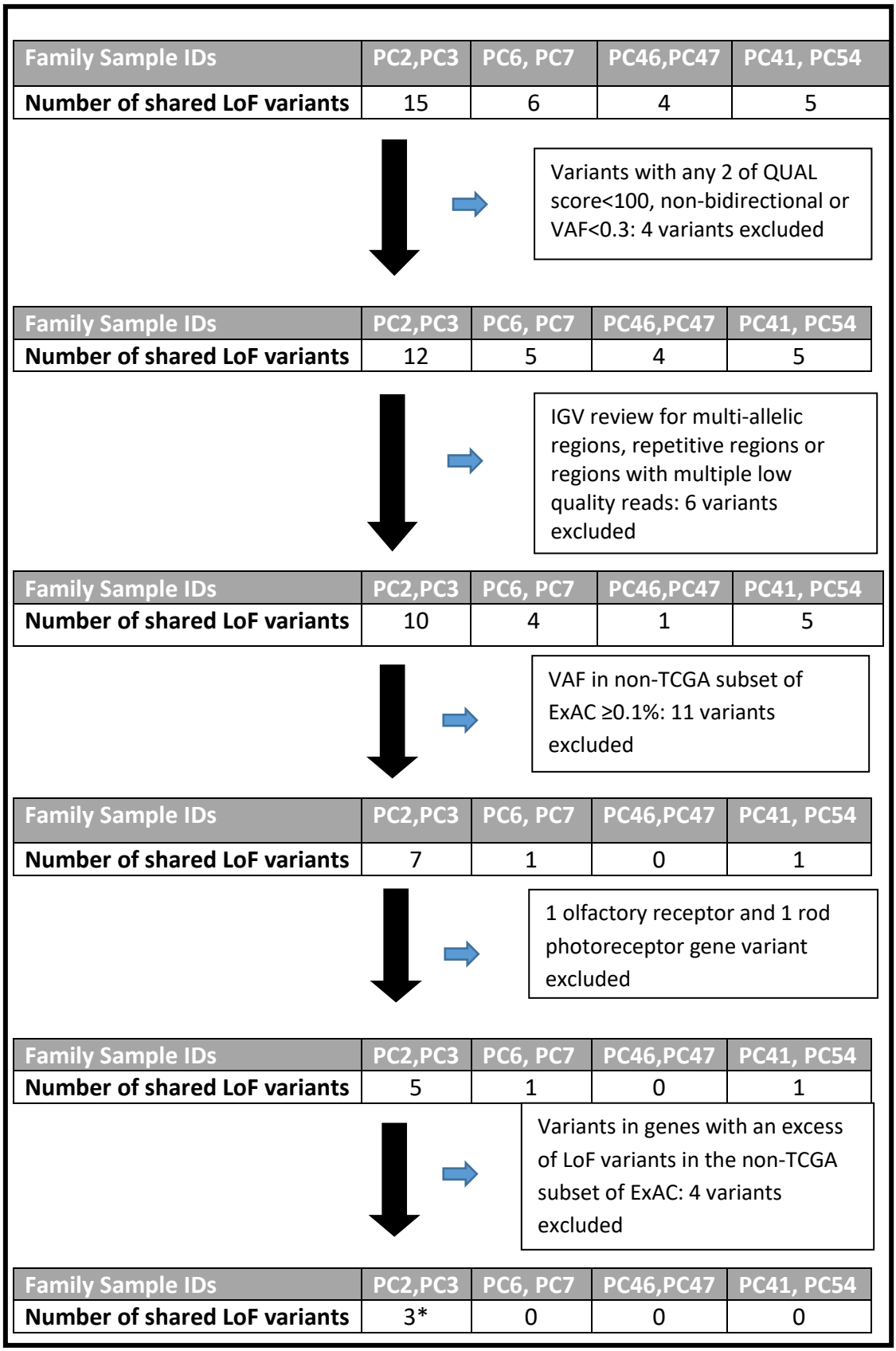



**Figure 4-6. Number of LoF variants seen in both family members at each step of the variant prioritisation process.**

LoF, loss-of-function. VAF, variant allele frequency. TCGA, The Cancer Genome Atlas. ExAC, Exome Aggregation Consortium. *APLF gene was not excluded, as discussed in the text.

LoF variants were omitted if any two of the following criteria were met: QUAL score <100, variant was not called off bidirectional reads, or the variant frequency was <0.3. All LoF variants were visualised in the Integrated Genome Viewer (Robinson et al., 2011). Variants from multi-allelic regions, repetitive regions or regions with multiple low quality reads were excluded.

In order to prioritise rarer variants, LoF variants with an allele frequency of <0.1% in the Non-TCGA Subset of ExAC were prioritised. An olfactory receptor and a rod photoreceptor gene were deprioritised due to implausible function and space constraints on the validation panel.

#### *4.3.5.1.2.2. Exclusion of genes with an excess of LoF variants in the Non-TCGA Subset of the ExAC dataset*

As the likelihood of FDRs and SDRs sharing a LoF variant is much higher than for 2 unrelated individuals, further evidence at a gene level was sought that such LoF variants may be pathogenic. Genes with LoF variants in a significant percentage of the population are unlikely to predispose to diseases that are rare in the population, particularly if the mutations are highly penetrant. Using known CRC-predisposing genes as a starting point, the percentage of individuals in the Non-TCGA subset of the ExAC database with a LoF variant in a known CRC-predisposing gene was first estimated. This estimate was used as a cut-off to de-prioritise novel genes with a LoF variant in a percentage of individuals exceeding that seen with the known CRC-predisposing genes. The following steps were performed: 1)The percentage of individuals with LoF variants in 26 228 genes was estimated from the Non-TCGA Subset of the ExAC dataset, 2) The estimated percentage of individuals in the Non-TCGA Subset of the ExAC dataset with LoF variants in the known CRC-predisposing genes was determined, and 3)a cut-off percentage was chosen based on the estimated percentage from 2), to de-prioritise genes with LoF variants in a higher percentage of the Non-TCGA subset of the ExAC dataset, as LoF variants in these genes are less likely to predispose to CRC, assuming a moderate to high penetrance.

The estimated percentage of individuals with a LoF variant in the Non-TCGA Subset of the ExAC dataset was determined as follows: Using the canonical transcript, the number of LoF alleles in the Non-TCGA Subset of the ExAC dataset was extracted by Richard Lupat from the Bioinformatics Core facility of PMCC.

Two assumptions were made in this assessment: 1) that each LoF variant was seen in a different individual and 2) that all 106210 alleles in the subset were sequenced. The percentage of individuals with a LoF variant was estimated by dividing the total number of LoF alleles by the total number of alleles in the Non-TCGA Subset of the ExAC dataset (106210), and multiplying by 200 (as each individual has 2 copies of an allele). This calculation gave a minimum estimate of the percentage of individuals with a LoF variant, as the sequence coverage is not 100% for every gene in every exome. We aimed to exclude genes with a clear excess of LoF variants in the Non-TCGA Subset of the ExAC dataset, hence a minimum estimate was acceptable. The minimum estimate of the percentage of individuals with a LoF variant in the known CRC-predisposing genes is shown in Table 4.12. For comparison, the same calculation was performed for *BRCA2*, a breast and ovarian cancer predisposing gene that has

a known common LoF variant (NP_000050.2:p.Lys3326Ter) in the last exon (Mazoyer et al., 1996).

**Table 4-12. Number of LoF alleles in known CRC-predisposing genes, and minimum estimate of percentage of individuals with a LoF variant, in the non-TCGA Subset of the ExAC dataset**

| Gene | No. of LoF alleles in Non-TCGA Subset of the ExAC dataset (canonical transcript) | Estimated percentage of individuals with a LoF variant in Non-TCGA Subset of the ExAC dataset |
|---|---|---|
| ***STK11*** | *0* | 0.00 |
| ***GREM1*** | *0* | 0.00 |
| ***SMAD4*** | *1* | 0.00 |
| ***MLH1*** | *8* | 0.02 |
| ***BMPR1A*** | *13* | 0.02 |
| ***APC*** | *14* | 0.03 |
| ***MSH2*** | *21* | 0.04 |
| ***EPCAM*** | *23* | 0.04 |
| ***MUTYH*** | *218* | 0.41 |
| ***PMS2*** | *408* | 0.77 |
| ***MSH6*** | *615* | 1.16 |
| ***BRCA2*** | 884 | 1.67 |

As the minimum estimate of the percentage of individuals with a LoF variant in the known CRC-predisposing genes was 1.16%, a cut-off of 2% or more was chosen to de-prioritise any genes that may be tolerant to LoF variants. The conservative 2% was chosen to accommodate genes that may have a commonly occurring LoF variant in the last exon, as is the case with *BRCA2*. LoF variants in the last exon of a gene generally do not result in nonsense-mediated mRNA decay (Chang et al., 2007, Perrin-Vidoz et al., 2002, Zhang et al., 1998a, Zhang et al., 1998b), with translation of a truncated protein that may still be functional. Therefore, the percentage of individuals with a LoF variant in the general population may be higher than expected, due to a percentage of individuals having a LoF variant in the last exon of a gene thath is not selected against.

Of the genes with shared LoF variants within the 4 families, 5 genes had a minimum estimated percentage of individuals with a LoF variant in the Non-TCGA Subset of the ExAC dataset of >2%. The LoF variants in these genes were inspected in the ExAC browser to determine if there was a common LoF variant in the last exon accounting for the majority of LoF variants in the gene, as is the case with *BRCA2*. For one gene (*APLF*), 1207 out of 1362 LoF variants in APLF consisted of a single frameshift variant in the last exon. *APLF* had a minimum estimated percentage of individuals with a LoF variant in the Non-TCGA Subset of the ExAC dataset of 2.1%. The LoF variant shared between PC2 and PC3 occurred in the first exon. *APLF* was included in the list of candidate CRC-predisposing genes. The other 4 genes did not have a common LoF variant in the last exon accounting for the majority of the LoF variants, and hence were excluded (Table 4.13).

**Table 4-13. Excluded genes with a shared LoF variant in a family**

| Gene | Family | No. of LoF alleles in Non-TCGA Subset of the ExAC dataset (canonical transcript) | Estimated percentage of individuals with a LoF variant in Non-TCGA Subset of the ExAC dataset |
|---|---|---|---|
| ***BCLAF1*** | PC2 and PC3 | 1735 | 3.27 |
| ***MROH2A*** | PC2 and PC3 | 2013 | 3.79 |
| ***ABCC11*** | PC6 and PC7 | 1390 | 2.62 |
| ***MUC19*** | PC41 and PC54 | 16943 | 31.90 |

Three shared LoF variants that passed all the preceding filters were identified from PC2 and PC3, the father and daughter with rectal adenocarcinoma (Table 4.14).

**Table 4-14. LoF variants identified in both father (PC3) and daughter (PC2) affected by rectal adenocarcinoma.**

| Gene | Gene Description | Ensembl Gene | Genomic Position | Ensembl Transcript | cDNA position | Minimum estimate of percentage of individuals with LoF variant in Non-TCGA Subset of ExAC dataset |
|---|---|---|---|---|---|---|
| ***NEDD4*** | neural precursor cell expressed, developmentally down-regulated 4, E3 ubiquitin protein ligase | ENSG00000069869 | 15:56216898C>T | ENST00000435532.3 | c.238-1G>A | 0.08 |
| ***TIMELESS*** | timeless circadian clock | ENSG00000111602 | 12:56811686_56811689delCCTA | ENST00000229201.4 | c.3553+2_3553+5delGGAT | 0.28 |
| ***APLF*** | aprataxin and PNKP like factor | DNA repair | ENSG00000169621 | 2:68694913delC | c.50delC | 2.10* |

*common LoF variant in the last exon

#### *4.3.5.1.2.3. Rare potentially pathogenic missense variants/in-frame indels (Consequence Rank 3) in both affected family members within a family*

Three of the four affected families consisted of two affected FDRs, and were hence expected to share fifty percent of their variants. The fourth family consisted of two affected SDRs and were expected to share twenty-five percent of their variants. The 3 pairs of affected FDRs shared 118, 100 and 100 potentially pathogenic Consequence Rank 3 variants, while the affected SDRs shared 50 potentially pathogenic Consequence Rank 3 variants.

In order to prioritise the missense variants most likely to be pathogenic, the rarest shared missense variants with the greatest predicted deleteriousness were chosen. Genes with private missense variants i.e. variants not seen in the Non-TCGA ExAC dataset, with a CADD Phred score >20 were prioritised (Table 4-15). One to seven variants were identified per family (Table 4.16). Of the 14 genes with shared intra-family missense variants, 1 gene harboured a LoF variant in a different family (Table 4.14), 2 genes harboured different potentially pathogenic missense variants in one additional family, while 1 gene harboured 2 different missense variants in two different families (Table 4.15).

**Table 4-15. Number of shared Consequence Rank 3 variants between family members, at each step of the filtering process**

| Intra-family sample IDs | Number of shared potentially pathogenic Consequence Rank 3 variants | Number of shared variants not seen in the Non-TCGA Subset of ExAC *i.e.* novel | Number of shared novel variants with a CADD scaled score>20 |
|---|---|---|---|
| **PC2, PC3** | 118 | 34 | 7 |
| **PC6, PC7** | 100 | 14 | 4 |
| **PC46, PC47** | 100 | 21 | 2 |
| **PC41, PC54** | 50 | 7 | 1 |

**Table 4-16. Private missense variants shared by both affected family members within a family, and LoF or missense variants in the same gene identified in unrelated individuals in the discovery cohort**

| Family | Gene | Gene Description | Ensembl Protein | Amino acid change | CADD scaled score | Rare LoF variant in a different family | Rare missense variant in a different family (CADD >10) | Same variant? | Sample | Ensembl Protein | Amino acid change | CADD scaled score | Frequency in Non-TCGA Subset of ExAC |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **PC2_3** | *INPP5E* | inositol polyphosphate-5-phosphatase, 72 kDa | ENSP00000360777.3 | p.Phe480Leu | 35.0 | 1 | . | . | . | ENSP00000360777.3 | p.Asn159Ter | . | 9.67E-06 |
| **PC2_3** | *GOLGA5* | golgin A5 | ENSP00000163416.2 | p.Asp681Asn | 34.0 | . | . | . | . | . | . | . | . |
| **PC2_3** | *PLSCR3* | phospholipid scramblase 3 | ENSP00000459019.1 | p.Pro259Leu | 33.0 | . | . | . | . | . | . | . | . |
| **PC2_3** | *HTR5A* | 5-hydroxytryptamine (serotonin) receptor 5A, G protein-coupled | ENSP00000287907.2 | p.Asp121Asn | 27.5 | . | . | . | . | . | . | . | . |
| **PC2_3** | *HCN3* | hyperpolarization activated cyclic nucleotide-gated potassium channel 3 | ENSP00000357342.3 | p.Cys311Ser | 24.5 | . | . | . | . | . | . | . | . |
| **PC2_3** | *GPR137B* | G protein-coupled receptor 137B | ENSP00000355551.3 | p.Glu150Gln | 24.0 | . | . | . | . | . | . | . | . |
| **PC2_3** | *SF3A3* | splicing factor 3a, subunit 3, 60kDa | ENSP00000362110.4 | p.Arg470Gln | 23.7 | . | . | . | . | . | . | . | . |
| **PC6_7** | *FKBP10* | FK506 binding protein 10, 65 kDa | ENSP00000317232.4 | p.Ser419Leu | 27.5 | . | . | . | . | . | . | . | . |
| **PC6_7** | *IL4R* | interleukin 4 receptor | ENSP00000379111.2 | p.Gly722Ser | 27.3 | . | . | . | . | . | . | . | . |
| **PC6_7** | *LRRK2* | leucine-rich repeat kinase 2 | ENSP00000298910.7 | p.Leu805His | 22.6 | . | 2 | No | PC39<br>PC3, PC4 | ENSP00000298910.7<br>ENSP00000298910.7 | p.Ile32Met<br>p.Arg2477Gln | 16.55<br>10.35 | .<br>8.48E-05 |
| **PC6_7** | *SUCLG2* | succinate-CoA ligase, GDP-forming, beta subunit | ENSP00000307432.5 | p.Leu349Phe | 20.7 | . | . | . | . | . | . | . | . |

**Table 4-16 (continued). Private missense variants shared by both affected family members within a family, and LoF or missense variants in the same gene identified in unrelated individuals in the discovery cohort**

| Family | Gene | Gene Description | Ensembl Protein | Amino acid change | CADD scaled score | Rare LoF variant in a different family | Rare missense variant in a different family (CADD >10) | Same variant ? | Sample | Ensembl Protein | Amino acid change | CADD scaled score | Frequency in Non-TCGA Subset of ExAC |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **PC46_47** | *UFL1* | UFM1-specific ligase 1 | ENSP00000358283.4 | p.Glu315Ala | 25.4 | . | 1 | No | PC24 | ENSP00000358283.4 | p.Arg191Cys | 17.08 | 7.25E-04 |
| **PC46_47** | *MMP2* | matrix metallopeptidase 2 (gelatinase A, 72kDa gelatinase, 72kDa type IV collagenase) | ENSP00000219070.4 | p.Lys531Glu | 21.4 | . | 1 | No | PC3 | ENSP00000219070.4 | p.Ala545Val | 19.56 | 9.42E-05 |
| **PC41_54** | *TMCC3* | transmembrane and coiled-coil domain family 3 | ENSP00000261226.4 | p.Lys121Arg | 30.0 | . | . | . | . | . | . | . | . |

### *4.3.5.1.2.4. Extensive analysis of a father and daughter with familial AMTs (manuscript submitted for publication)*

**Germline whole exome sequencing of a family with appendiceal mucinous tumours presenting with pseudomyxoma peritonei**

Mei Sim Lung[1], Catherine A. Mitchell[2], Maria A. Doyle[3], Andrew C. Lynch[4], Kylie L. Gorringe[5,7], David D.L. Bowtell [6,7,8], Australian Ovarian Cancer Study Group, Ian G. Campbell[1,2,7¶], Alison H. Trainer[1,9,10¶]

[1]Cancer Genetics Laboratory, Peter MacCallum Cancer Centre, Melbourne, Victoria, Australia

[2]Department of Pathology, Peter MacCallum Cancer Centre, Melbourne, Victoria, Australia

[3]Research Computing Facility, Peter MacCallum Cancer Centre, Melbourne, Victoria, Australia

[4]Department of Surgical Oncology, Peter MacCallum Cancer Centre, Melbourne, Victoria, Australia

[5] Cancer Genomics Program, Peter MacCallum Cancer Centre, Melbourne, Victoria, Australia

[6]Cancer Genetics and Genomics Laboratory, Peter MacCallum Cancer Centre, Melbourne, Victoria, Australia

[7]Sir Peter MacCallum Department of Oncology, University of Melbourne, Parkville, Victoria, Australia

[8]Department of Biochemistry and Molecular Biology, University of Melbourne, Parkville, Victoria, Australia

[9] Parkville Familial Cancer Centre, Peter MacCallum Cancer Centre, Melbourne, Victoria, Australia

[10] Department of Medicine, University of Melbourne, Parkville, Victoria, Australia

¶ co-senior authors

**Corresponding author**

Mei Sim Lung

Research Division

Peter MacCallum Cancer Centre

Victorian Comprehensive Cancer Centre Building

305 Grattan St

Melbourne VIC 3000

Australia

Tel no: (+613) 85596518

Fax no: (+613) 8559 7739

E-mail address:mei.lung@petermac.org

**Abstract**

Background

Familial cases of appendiceal mucinous tumours (AMTs) are extremely rare and the underlying genetic aetiology uncertain. We identified potential predisposing germline genetic variants in a father and daughter with AMTs presenting with pseudomyxoma peritonei (PMP) and correlated these with regions of loss of heterozygosity (LOH) in the tumours.

Materials and Methods

Through germline whole exome sequencing, we identified novel heterozygous loss-of-function (LoF) (i.e. nonsense, frameshift and essential splice site mutations) and missense variants shared between father and daughter, and validated all LoF variants, and missense variants with a Combined Annotation Dependent Depletion (CADD) scaled score of ≥10. Genome-wide copy number analysis was performed on tumour tissue from both individuals to identify regions of LOH.

Results

Seventeen novel variants in 17 genes were shared by the father and daughter: a nonsense mutation in *REEP5*, an essential splice site mutation in *THOP1*, and 15 missense variants. None of these germline variants were located in tumour regions of LOH shared by the father and daughter. Four genes (*EXOG*, *RANBP2, RANBP6 and TNFRSF1B*) harboured missense variants that fell in a region of LOH in the tumour from the father only, but none showed somatic loss of the wild type allele in the tumour. The *REEP5* gene was sequenced in 23 individuals with presumed sporadic PMP; no LoF or rare missense germline variants were identified.

Conclusion

Germline exome sequencing of a father and daughter with AMTs identified novel candidate predisposing genes. Further studies are required to clarify the role of these genes in familial AMTs.

Keywords: pseudomyxoma peritonei, appendiceal tumour, familial, germline predisposition, exome sequencing

**Background**

Appendiceal mucinous tumours (AMTs) are rare, occurring at an age-adjusted incidence of 0.12 per million individuals, with a median age at diagnosis of 59 years and no gender bias (1). Pseudomyxoma peritonei (PMP) is a clinical term describing gelatinous ascites, associated with the presence of mucin-producing cells within the peritoneal cavity, usually associated with an appendiceal mucinous neoplasm, either a low-grade appendiceal mucinous neoplasm (LAMN) or a mucinous adenocarcinoma (2). Whether PMP occurs in all, or just a molecular subset of AMTs is unknown.

Molecular studies of PMP are limited, but somatic mutations have been identified in *KRAS in* 60-100% of tumours, in *GNAS* in 30-69% of tumours and in *TP53* in ~14% of tumours (3-8). The reported frequency of somatic mutations in the *APC* gene varies widely between studies ranging between 0-33% (4, 7).

Familial forms of AMTs are rare, with only two reported cases in the literature. The first family comprised monozygotic twin brothers (9). The first twin was diagnosed with PMP at the age of 35 during an umbilical hernia repair, and he was subsequently found to have a perforated AMT. After this diagnosis, his asymptomatic twin underwent a prophylactic appendectomy, which identified a non-perforated AMT. Somatic LOH of the *APC* locus was identified in the second AMT but not in the original case. No germline *APC* mutation data were available.

The second family comprised a brother and sister diagnosed with AMTs at the ages of 69 and 77 years respectively (10). The brother presented with acute appendicitis and an AMT was identified at surgery, while his sister presented with increasing abdominal girth and was found to have PMP and an AMT. An assessment for Lynch syndrome was performed in this family. The sister's AMT had normal immunohistochemistry staining for the mismatch repair proteins MLH1, MSH2, MSH6 and PMS2, and her tumour was microsatellite stable, whilst her brother

underwent constitutional mutation analysis of *MLH1, MSH2* and *MSH6,* which did not identify a pathogenic variant.

Here, we report the first familial parent-child PMP case, a father (P1) and daughter (P2) who were both diagnosed with PMP secondary to an AMT at the ages of 66 and 51 years respectively. P1 was incidentally found to have PMP on staging CT for a Gleason 6 prostate cancer. A diagnosis of PMP was made on diagnostic laparotomy and he underwent drainage of a large abdominal cyst. Subsequent cytoreductive surgery undertaken 4 years later when he became symptomatic, identified a ruptured LAMN. His daughter (P2) presented with 6 months of menorrhagia, dysmenorhoea and pelvic and right upper quadrant discomfort. She was found to have PMP and a moderately differentiated mucinous adenocarcinoma of the appendix.

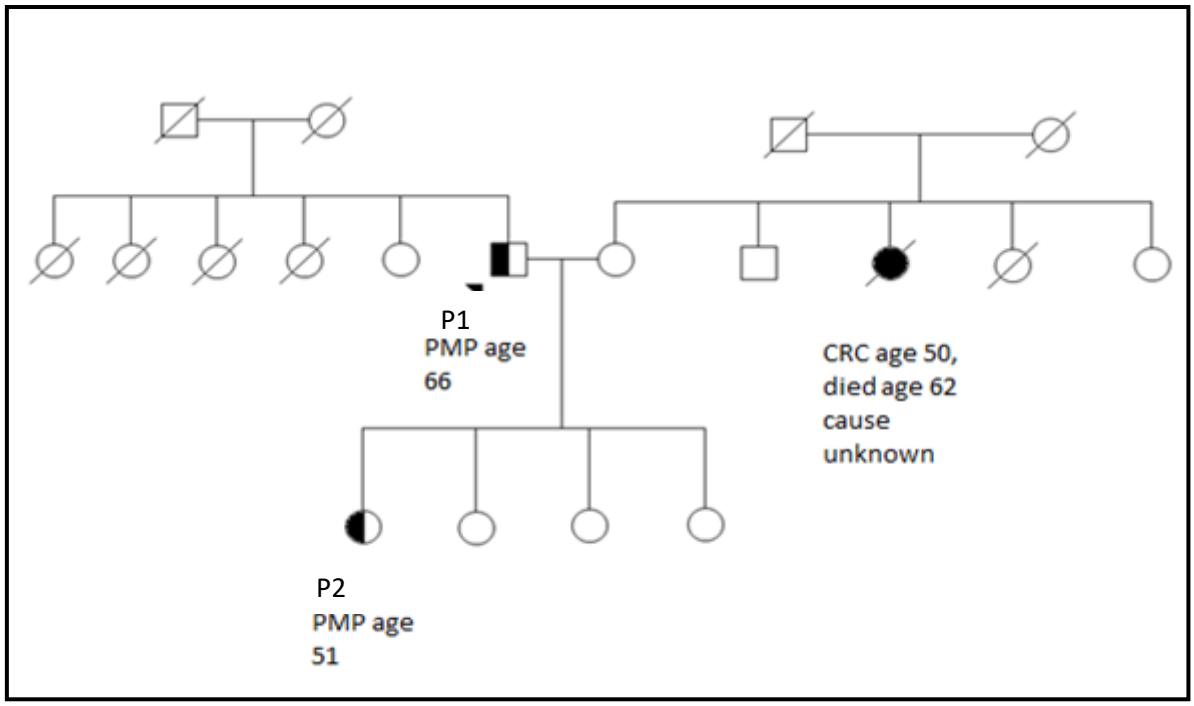



**Figure 1.** Pedigree of parent child pseudomyxoma peritonei (PMP). The father (P1) presented with PMP at age 66, while his daughter (P2) developed PMP at age 51.

Identifying the germline genetic aetiology of rare familial colorectal cancer syndromes such as familial adenomatous polyposis has led to a better understanding of many somatic pathways and mechanisms underlying sporadic forms of the disease. Thus, we sought to identify predisposing genes in this family through a germline whole exome sequencing approach and looked for regions of LOH in their tumour tissues as a secondary filter for potentially pathogenic variants.

## Materials and Methods

### Ethics

All procedures performed involving human participants were in accordance with the ethical standards of the Peter MacCallum Cancer Centre Human Research and Ethics Committee (project number 10_83) and with the 1964 Helsinki declaration and its later amendments or comparable ethical standards. Informed consent was obtained from all individual participants included in the study.

### Whole exome sequencing

One µg of germline DNA was obtained from peripheral leucocytes and fragmented using the Covaris S2 System (Covaris, Woburn, MA, USA). The SureSelect Human All Exon v1 (Agilent, Santa Clara, CA, USA) was used for exome enrichment according to the manufacturer's protocol. Paired-end 100 base pair reads were sequenced on a HiSeq2000 (Illumina Inc, San Diego, CA, USA) instrument. Both exomes passed sequencing quality control with mean target base coverages of 129x and 121x for P1 and P2 respectively and >95% of targeted bases covered more than 10x.

### Sequence alignment and variant calling

Raw sequence reads were quality checked with FastQC (11) and trimmed for low quality bases and adaptor if necessary using Cutadapt (12). Reads were aligned to the human genome (GRCh37 assembly) using BWA-MEM (13). Duplicate reads were marked using Picard (14) followed by merging of BAM files for both individuals. Local realignment around indels was performed on the merged BAM files using the Genome Analysis Tool Kit (GATK) software v3.1 (15). Subsequently, base quality score recalibration was performed using GATK software. Single nucleotide variants (SNVs) and indels were identified using the GATK HaplotypeCaller

and annotated with information from Ensembl release 73 using Ensembl's Perl API and Variant Effect Predictor (16, 17). Each variant was annotated with its frequency in the 1000 Genomes Project (18), the National Heart, Lung and Blood Institute (NHLBI) Grand Opportunity (GO) Exome Sequencing Project (19) and an in-house exome dataset of 147 familial breast cancer cases (20). The likely pathogenic consequence for each variant was determined by Polyphen (21), SIFT (22), and Combined Annotation Dependent Depletion (CADD) scaled score (23).

**Exome data analysis**

For genes with multiple transcripts, transcripts were prioritised on 1) most to least deleterious predicted impact of variant on protein function (Supplementary Data 1), and 2) RefSeq transcript. The highest ranking transcript was taken forward for further analysis. LoF variants and missense variants that met the following criteria were considered for further analysis: (1) Phred variant quality score of >30, and (2) variant allele frequency between 0.15 and 0.8. For identification of novel variants shared between P1 and P2, variants were excluded if they were present in control cohorts: 1000 Genomes Project, NHLBI GO Exome Sequencing Project or an in-house cohort of 147 Australian familial breast cancer exomes. All loss of function (LoF) variants (truncating frameshift, nonsense, essential splice site), and missense variants with a CADD scaled score ≥10 were manually checked in the Integrated Genome Viewer (IGV) (24, 25).

Variants shared between P1 and P2 that were confirmed on Sanger sequencing were checked in the Exome Aggregation Consortium (ExAC) dataset, a dataset comprising exome data from 60,706 unrelated individuals (26), for the population frequency, to confirm that these variants were rare or novel.

**Whole genome amplification and Sanger sequencing**

Candidate variants were confirmed by Sanger sequencing using whole-genome amplified DNA from P2. Whole-genome amplification of genomic DNA was performed using the REPLI-g Midi Kit (Qiagen, Redwood City, CA, USA). PCR primers were designed using the Primer3 program v0.4.0 (27, 28) and are listed in Supplementary Data 2. DNA fragments were amplified using HotStarTaq DNA Polymerase (Qiagen, Redwood City, CA, USA), purified using ExoSAP-IT PCR Purification Kit (USB Corporation, Cleveland, OH, USA), and sequenced using the Big Dye Terminator v3.1 kit (Applied Biosystems, Foster City, CA, USA). Sanger sequencing was performed on an ABI 3130 Sequencer (Applied Biosystems), and visualised in Geneious 5.6.2 software (BioMatters Ltd, Auckland, New Zealand).

**Tumour micro-dissection and analysis**

Both tumours were reviewed by a clinical pathologist with expertise in this area. Consecutive 10 µm sections were cut from the formalin fixed paraffin embedded PMP specimens with the highest tumour content, and stained with haematoxylin and eosin. Tumour cells were micro-dissected manually using a 23 gauge needle and somatic DNA was extracted using the DNeasy Blood and Tissue Kit (Qiagen, Redwood City, CA, USA). Somatic copy number analysis of tumours was assayed using the OncoScan Molecular Inversion Probe assay (*Affymetrix,* Santa Clara, CA, USA) on 50-75 ng of somatic DNA, and the data analysed using Nexus Copy Number™ software (Biodiscovery, Inc., El Segundo, CA, USA). There was no matched control copy number data available for P2. The Oncoscan molecular assay comprises >220,000 single nucleotide polymorphisms and provides copy number resolution around 50-100 kb.
To assess if a variant showed somatic LOH, Sanger sequencing was performed using unamplified tumour DNA extracted from the AMT (primers listed in Supplementary Data 3).

**Sanger sequencing of candidate genes in a PMP validation cohort**

Germline DNA from individuals with sporadic PMP was obtained from the Victorian Cancer Biobank (VCB), the Australian Ovarian Cancer Study (AOCS) and Southampton, UK (29). Histopathology reports for all PMP samples in the validation cohort were examined to ensure that they were not metastases of ovarian origin.

Sanger sequencing of all exons of *REEP5* was performed using germline DNA from individuals from the PMP validation cohort, using the same methods as described earlier. PCR primers for each exon were designed to include 40 base pairs flanking the intron-exon boundary of each exon (Supplementary Data 4).

## Results

### Analysis of genes associated with known familial colorectal cancer syndromes

Prior to analysis for novel shared germline variants, a targeted analysis was undertaken of 14 genes associated with an increased familial colorectal cancer risk*: APC, MUTYH, MLH1, MSH2, MSH6, PMS2, BMPR1A, SMAD4, STK11, EPCAM (30), GREM1, POLE (31), POLD1 (31) and NTHL1 (32).* All these genes excluding *STK11* were sequenced with a mean coverage of > 68X (Supplementary Data 5). No LoF or known pathogenic missense variants were identified in either case. A rare missense variant in *POLE* carried by both P1 and P2 was identified*, POLE* p.Val1887Met (NM_006231.2:c.5659C>T). This variant has an allele frequency of $6.15X10^{-4}$ in the NHLBI GO Exome Sequencing Project. This *POLE* variant did not fall in a region of LOH in either P1 or P2. Using in-silico prediction tools, this variant had a CADD score of 10.07, but was predicted non-pathogenic by both SIFT and PolyPhen.

### Identification of shared germline variants

LoF and missense variants with an allele frequency between 0.15 and 0.8 were selected in order to identify heterozygous variants. A total of 4,893 and 4,973 variants were identified in P1 and P2, respectively. After excluding any variant previously reported in the 1000 Genomes Project or the NHLBI GO Exome Sequencing Project, or a local cohort of 147 familial breast cancer exomes, a total of 106 and 110 variants remained in P1 and P2, respectively. Of these, 40 variants (8 LoF and 32 missense) were shared between P1 and P2. Manual curation of the variant reads in IGV eliminated a further 9 variants as likely artefacts due to misalignment, known hypervariable genes or variants located in areas with low mapping quality. Missense variants were prioritised on their CADD score. A CADD scaled score of ≥10 was used as an inclusion threshold, as this identifies the top 10% most deleterious substitutions in the human genome (23). This approach eliminated a further seven missense variants. Following validation

by Sanger sequencing, the final list of shared variants in 17 genes comprised 2 LoF and 15 missense variants (Table 1). The number of variants remaining after each filtering step is summarised in Supplementary Data 6.

**Table 1. Seventeen validated germline variants seen in both P1 and P2**

| Gene | Ref_Seq mRNA transcript | cDNA variant | Amino Acid Change | Variant type | Scaled CADD score | ExAC variant frequency | Located in LOH region in tumour (sample ID) |
|---|---|---|---|---|---|---|---|
| ***REEP5*** | NM_005669.4 | c.159T>G | p.Tyr53* | Nonsense | 31 | 0 | No |
| ***THOP1*** | NM_003249.3 | c.1895_2025del | - | Essential splice site | 12.83 | 0 | No |
| ***RHBDL2*** | NM_017821.3 | c.395C>G | p.Gly132Ala | Missense | 29.8 | 0 | No |
| ***FGFR4*** | NM_022963.2 | c.1988G>A | p.Arg663Gln | Missense | 21.9 | 8.87E-06 | No |
| ***CNTN2*** | NM_005076.3 | c.2263A>T | p.Ser755Cys | Missense | 21.2 | 8.27E-06 | No |
| ***RANBP2*** | NM_006267.4 | c.3683G>T | p.Gly1228Val | Missense | 19.8 | 0 | Yes (P1) |
| ***HSPB6*** | NM_144617.2 | c.59G>A | p.Pro20Leu | Missense | 19.22 | 0.0006623 | No |
| ***ZNF747*** | NM_023931.2 | c.254C>T | p.Gly85Glu | Missense | 16.28 | 0 | No |
| ***EXOG*** | NM_005107.3 | c.178G>A | p.Ala60Thr | Missense | 16.17 | 0 | Yes (P1) |
| ***BRINP3*** | NM_199051.1 | c.809A>C | p.Glu270Ala | Missense | 16.1 | 8.284e-06 | No |
| ***ASIC1*** | NM_020039.3 | c.463C>T | p.Arg155Cys | Missense | 16.05 | Not covered | No |
| ***TNFRSF1B*** | NM_001066.2 | c.1337A>G | p.Glu446Gly | Missense | 15.46 | Not covered | Yes (P1) |
| ***RANBP6*** | NM_001243202.1 | c.251A>C | p.Glu84Ala | Missense | 14.87 | 2.64e-05 | Yes (P1) |
| ***LSR*** | NM_205834.3 | c.725C>T | p.Thr242Ile | Missense | 14.79 | 2.64E-05 | No |
| ***MTERFD3*** | NM_001033050.1 | c.607C>T | p.Ala203Thr | Missense | 12.04 | 1.66E-05 | No |
| ***PAX1*** | NM_001257096.1 | c.1379G>T | p.Arg460Leu | Missense | 11.57 | 8.48E-05 | No |
| ***EGFR*** | NM_005228.3 | c.1915A>C | p.ASN639His | Missense | 10.51 | 0 | No |

CADD,Combined Annotation Dependent Depletion Score. ExAC, Exome Aggregation Consortium. Sample ID refers to tumour sample identification.

**Identification of germline variants associated with concomitant areas of tumour LOH**

Assuming a possible two-hit model for germline tumour suppressor inactivation, analysis was performed to identify germline candidate variants present in regions of LOH in the tumours. The Oncoscan array detected somatic LOH (copy number neutral and loss) in 20.3% and 19.8% of the tumour genomes of P1 and P2, respectively (Tables 2 and 3). Only one region of LOH, a 1.77 Mb region on chromosome 2q, was common to both tumours (Fig. 3). This region contains 33 genes but none harboured likely pathogenic germline variants common to both individuals.

**Table 2. Regions of loss of heterozygosity (LOH) and copy number (CN) loss in pseudomyxoma peritonei tumour from P1**

| Chromosome | Start | End | Event | Gene containing shared germline variant |
|---|---|---|---|---|
| 1 | 0 | 26948921 | CN Loss/LOH | *TNFRSF1B* |
| 1 | 147134028 | 147825662 | LOH | - |
| 1 | 190881536 | 191887248 | LOH | - |
| 2 | 50914938 | 51446707 | LOH | - |
| 2 | 99377108 | 243199373 | LOH | *RANBP2* |
| 3 | 0 | 48040095 | LOH | *EXOG* |
| 3 | 50384337 | 52768237 | LOH | - |
| 3 | 191022128 | 191856841 | LOH | - |
| 4 | 21562094 | 22526745 | LOH | - |
| 6 | 33506076 | 171115067 | CN Loss/LOH | - |
| 7 | 22818202 | 23530679 | LOH | - |
| 9 | 0 | 33946637 | LOH | *RANBP6* |
| 9 | 118867206 | 119530524 | LOH | - |
| 10 | 62401757 | 63554609 | LOH | - |
| 10 | 69912047 | 70990019 | LOH | - |
| 11 | 5529179 | 6091608 | LOH | - |
| 11 | 34356991 | 35043999 | LOH | - |
| 12 | 21011988 | 22928787 | LOH | - |
| 12 | 29116987 | 29816788 | LOH | - |
| 12 | 111166777 | 112632998 | LOH | - |
| 14 | 36158966 | 36786121 | LOH | - |
| 15 | 84783628 | 85501061 | LOH | - |
| 17 | 857820 | 1452131 | LOH | - |
| 18 | 18535946 | 19513726 | LOH | - |
| 19 | 41994722 | 48382347 | CN Loss/LOH | - |
| 21 | 14414872 | 48129895 | CN Loss/LOH | - |

CN, copy number. LOH, loss of heterozygosity.

**Table 3. Regions of loss of heterozygosity (LOH) and copy number (CN) loss in pseudomyxoma peritonei tumour from P2**

| Chromosome | Start | End | Event |
|---|---|---|---|
| 2 | 241428066 | 243199373 | CN Loss/LOH |
| 8 | 0 | 33395041 | CN Loss/LOH |
| 8 | 42137047 | 49106261 | LOH |
| 10 | 0 | 29922641 | CN Loss |
| 11 | 66296148 | 135006516 | LOH |
| 11 | 111333148 | 117815955 | CN Loss |
| 11 | 130316743 | 135006516 | CN Loss |
| X | 1 | 155270560 | CN Loss/LOH |

CN, copy number. LOH, loss of heterozygosity.

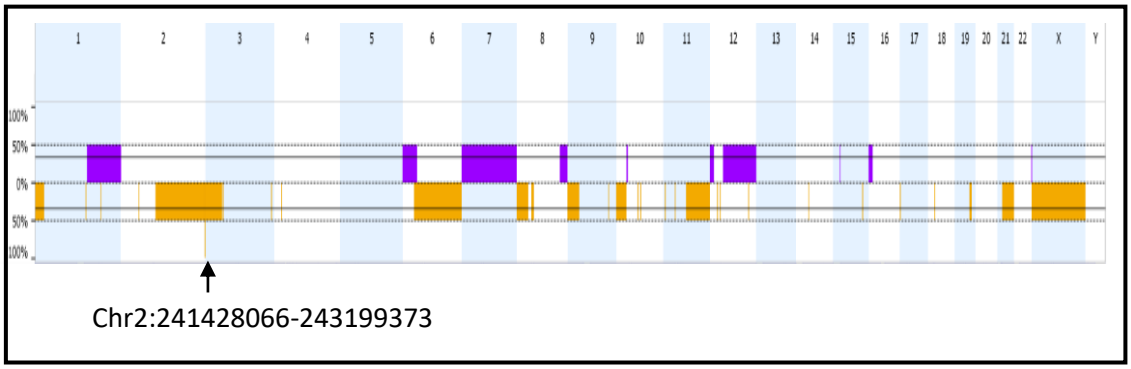



**Figure 2.** Summary plot of combined regions of copy number gain (blue) or loss of heterozygosity (brown), with bars reaching 100% indicating copy number aberrations seen in both tumours. The region on chromosome 2 containing loss of heterozygosity in both tumours is indicated by the arrow, with genomic coordinates given below. Chr, chromosome.

As LOH is not the only mechanism by which a somatic wild type allele can be abrogated, the analysis was extended to incorporate all regions of somatic LOH present in either P1 or P2's tumour. Missense variants in four genes, *EXOG*, *RANBP2, RANBP6 and TNFRSF1B*, were

identified in areas of LOH seen in the tumour from P1 (Table 1 and Table 2). The LOH regions in the tumour from P2 did not harbour any shared germline variants.

The variants in *EXOG*, *RANBP2, RANBP6 and TNFRSF1B* were sequenced in the tumour DNA of P1 to assess which allele of the gene was lost. *RANBP2* showed somatic loss of the mutant allele. Unexpectedly, somatic sequencing of *EXOG*, *RANBP6* and *TNFRSF1B* demonstrated the presence of both the wild type and mutant alleles (Supplementary Data 7 and 8). These discordant results may relate to the genomic heterogeneity within subclones of the PMP tumours. The DNA used as template for Sanger sequencing was extracted from the primary appendiceal tumour whilst the template DNA for the Oncogene array was extracted from secondary peritoneal tumour in order to obtain sufficient DNA. Due to the sparse nature of the tumour cells, no remaining DNA was available for Sanger sequencing from the secondary peritoneal tumour.

**Screening of the *REEP5* gene in a validation cohort of individuals with sporadic PMP using Sanger sequencing**

Having identified a LOF variant in *REEP5* present in both P1 and P2, further independent evidence implicating this gene in AMT predisposition was found in a previous case report (33) of an individual with a chromosomal translocation t(5;8)(q22;p23.1) who was diagnosed with a mucin-secreting appendiceal carcinoma and familial adenomatous polyposis (FAP) at the age of 26 years. Fluorescent in situ hybridisation studies in this case indentified a microdeletion encompassing both *APC* and *MCC* genes, and by implication the intervening gene *REEP5* (also known as DP1). On the basis of this additional information, Sanger sequencing of the five exons of *REEP5 (*NM_005669.4) was performed on the germline DNA of 23 individuals with presumed sporadic PMP (12 male, 11 female). The age of diagnosis (extracted from the histopathology report) ranged from 31 to 72 years, with a median age of 58 years. The age for

one individual was unknown. No LoF variants or missense variants (excluding common polymorphisms) in *REEP5* were identified.

**Discussion**

As PMP is a rare disorder, we postulated that this apparent familial occurrence of the disease in a father and daughter might imply a hereditary predisposition. Only a limited assessment of the known colorectal cancer predisposing genes had been performed previously on familial PMP cases in the literature. Through germline whole exome sequencing of both affected individuals in our parent-child PMP family, we were able to assess and exclude the known colorectal cancer predisposing genes including *APC, MLH1, MSH2, MSH6, PMS2, MUTYH, BMPR1A, SMAD4, POLE and POLD1*, with the caveat that *STK11* could not be excluded due to low coverage on exome sequencing. The functional consequence of the *POLE* p.Val1887Met (NM_006231.2:c.5659C>T) variant is unknown and requires further study.

We took an agnostic approach to identifying the causative gene by identifying novel (ie. not present in the 1000 Genomes Project and NHLBI GO Exome Sequencing Project datasets) shared germline variants in the exome of both individuals, coupled with genome-wide copy number analysis of both tumours. We identified 17 potentially pathogenic variants shared between the two cases, including four within regions of LOH in the father's tumour. Loss of the potentially pathogenic missense variant in *RANBP2* in the tumour genome, suggests this variant is unlikely to be the cause of the PMP predisposition. The missense variants in *EXOG*, *RANBP6* and *TNFRSF1B* remained heterozygous in the primary tumour of P1, although with the loss of the second allele at a later stage, this does not necessarily exclude these variants as possible PMP predisposition genes on these data alone. The regions of LOH containing the shared variants in these 3 genes may have arisen in a subclone of the primary appendiceal tumour that subsequently metastasised to the peritoneum.

A shared nonsense variant in *REEP5* p.Tyr53* *(*NM_005669.4:c.159T>G) would appear to be a strong candidate as the predisposing variant, although both tumours retained heterozygosity

at this locus. *REEP5* has been implicated in the regulation of *TP53*, a known cancer predisposition gene, through its interaction with *HCCR1 (34, 35)*. It is expressed in normal colonic tissue and has been shown to be down-regulated in colon cancers. Similarly, transfection of *REEP5* into RKO colon cancer cells results in growth inhibition and induction of apoptosis, suggesting it functions as a tumour suppressor (34). *REEP5* is also involved in stabilisation of the endoplasmic reticulum tubules via its effect on the curvature of the endoplasmic reticulum lipid bilayer (36). The lack of *REEP5* mutations in the 23 sporadic PMP cases that were screened suggests that germline *REEP5* mutations are not common in sporadic PMP.

*THOP1,* which harboured a shared essential splice site variant, is an oligopeptidase involved in metabolism of neuropeptides such as bradykinin (37). Increased *THOP1* expression in background liver is associated with improved survival post resection of hepatocellular carcinoma (38), suggesting a possible tumour suppressive effect.

Several genes that contained a shared rare missense variant have functions consistent with cancer predisposition. *RHBDL2* is a rhomboid intra-membrane protease that activates epidermal growth factor and is associated with anoikis resistance (39). *FGFR4* is part of the fibroblast growth factor receptor family, a subset of tyrosine kinase receptors that are highly conserved. The fibroblast growth factor receptors have been extensively investigated somatically in relation to various cancers, and several fibroblast growth factor receptor inhibitors are in clinical trials (40). The *FGFR4* p.Arg663Gln (NM_022963.2:c.1988G>A) variant found in this family is only seen in 1 person in the ExAC dataset. The functional consequence of this variant is unknown, however it has a high CADD scaled score and is predicted to be deleterious by SIFT and potentially damaging by PolyPhen, making it an interesting candidate for further functional study. A common germline polymorphism (*FGFR4:*c.1162G>A) has been

associated with an increased risk of developing breast and prostate cancers (41). *EGFR* is a receptor tyrosine kinase that is somatically mutated in 10% of non-small cell lung cancers (42). Exon 19 deletions and a p.L858R mutation account for approximately 90% of such mutations, and predict exquisite sensitivity to EGFR tyrosine kinase inhibitors (45-47). The identified missense variant in this study lies outside the protein kinase domain, hence a potential role in the development of AMTs remains to be determined.

**Study limitations**

This study only searched exon-based sequences and as such would be unable to identify sequence variants present in gene regulatory regions. The aggregation of PMP within the family may also have arisen due to a shared environmental factor, or due to stochastic events. PMP is a hypo-cellular tumour with only widely scattered groups of tumour cells in abundant mucin. This phenotype provided technical problems in the somatic LOH analysis.

**Conclusions**

Through germline whole exome sequencing of a very rare familial occurrence of a very rare cancer type, in which only 2 families with limited genomic analyses have been reported in the literature, our study makes several contributions to knowledge.

We have identified candidate variants that may predispose to the development of AMTs and PMP in this family. We were also able to examine the known CRC-predisposing genes (apart from *STK11)* and found no pathogenic variants in this family, suggesting that other novel genes may predispose to this rare subtype of cancer in the colon. Through the re-sequencing of

*REEP5* in 23 sporadic PMP cases, we found that *REEP5* mutations are not a common cause of sporadic PMP despite a shared LoF variant found in our family, and the loss of *REEP5* in a case report of a patient with FAP and an AMT. Further studies are required to ascertain if the genes identified in this study play a role in the development of PMP in familial or sporadic cases.

**List of abbreviations**

| Abbreviation | Definition |
|---|---|
| **AMT** | Appendiceal mucinous tumour |
| **PMP** | Pseudomyxoma peritonei |
| **LOH** | Loss of heterozygosity |
| **LoF** | Loss-of-function |
| **CADD** | Combined Annotation Dependent Depletion |
| **LAMN** | Low-grade appendiceal mucinous neoplasm |
| **DNA** | Deoxyribonucleic acid |
| **GATK** | Genome Analysis Tool Kit |
| **SNV** | Single nucleotide variant |
| **NHLBI GO** | National Heart, Lung and Blood Institute Grand Opportunity |
| **ExAC** | Exome Aggregation Consortium |
| **PCR** | Polymerase chain reaction |
| **FAP** | Familial adenomatous polyposis |

**Supplementary Data**

Supplementary Data 1 Ranking of gene transcript according to predicted functional consequence of a variant.pdf

Supplementary Data 2 PCR primers for validation of loss-of-function (LoF) variants and missense variants with Combined Annotation Dependent Depletion scaled score ≥10.pdf

Supplementary Data 3 M13 PCR primers for determination of loss-of-heterozygosity (LoH) in tumour tissue.pdf

Supplementary Data 4 Primers for *REEP5* exons.pdf

Supplementary Data 5 Coverage of familial colorectal cancer genes in whole exome sequencing.pdf

Supplementary Data 6 Number of variants at each step of the filtering process.pdf

Supplementary Data 7 Sanger sequencing traces for the *RANBP2* missense variant in the germline of PC1 (top left), germline of PC2 (middle left), and tumour of PC1 (bottom left), showing loss of the mutant allele.pdf

Supplementary Data 8 Sanger sequencing traces for missense variants in *EXOG* (top panel), *RANBP6* (middle panel) and *TNFRSF1B* (bottom panel) showing the presence of the variant in the germline of PC1, PC2 and in tumour from PC1, despite the presence of loss of heterozygosity of approximately a) 50 Mbs b) 30Mbs and c)loss of heterozygosity and copy number loss of approximately 20 Mbs.pdf

**Declarations**

Ethics approval and consent to participate

All procedures performed involving human participants were in accordance with the ethical standards of the Peter MacCallum Cancer Centre Human Research and Ethics Committee (project number 10_83) and with the 1964 Helsinki declaration and its later amendments or comparable ethical standards. Informed consent was obtained from all individual participants included in the study.

## Consent to publication

P1 and P2 signed patient information and consent forms to participate in this study, which includes the consent for publication of de-identified data.

## Availability of data and materials

The datasets used and/or analysed during the current study are available from the corresponding author on reasonable request.

## Competing interests

The authors declare that they have no competing interests.

## Funding

This study was funded through a grant from the Australian National Health and Medical Research Council and Victorian Cancer Council (APP1028374). Mei Sim Lung is funded by the Australian Postgraduate Award through the University of Melbourne, Australia.

The Australian Ovarian Cancer Study (AOCS) was supported by the U.S. Army Medical Research and Materiel Command under **DAMD17**-01-1-0729, The Cancer Council Victoria, Queensland Cancer Fund, The Cancer Council New South Wales, The Cancer Council South Australia, The Cancer Foundation of Western Australia, The Cancer Council Tasmania and the National Health

and Medical Research Council of Australia (NHMRC; ID400413, ID400281), Ovarian Cancer Australia (OCA) and the Peter MacCallum Foundation.

## Author's contributions

MSL planned the study, performed whole genome amplification, Sanger sequencing and tumour microdissection, analysed and interpreted the whole exome sequencing data and tumour data and drafted the manuscript. CAM performed the histological examinations of the AMTs and PMP specimens. MAD performed sequence alignment and variant calling. ACL identified the father and daughter with AMTs. KLG edited the manuscript. DDLB is the principal investigator on the Australian Ovarian Cancer Study Group that provided the sporadic PMP specimens. IGC and AHT conceived and initiated the study, obtained funding, contributed to the interpretation of study data and edited the manuscript.

## Acknowledgements

The Molecular Core facility at Peter MacCallum Cancer Centre performed whole exome sequencing on germline DNA from P1 and P2. Biospecimens used in this research were also obtained from the Victorian Cancer Biobank, Victoria, Australia with appropriate ethics approval. The Victorian Cancer Biobank is supported by the Victorian Government.

**References**

**1.Shaib WL, Goodman M, Chen Z, Kim S, Brutcher E, Bekaii-Saab T, et al. Incidence and Survival of Appendiceal Mucinous Neoplasms: A SEER Analysis. American journal of clinical oncology. 2015 Aug 11. PubMed PMID: 26270447. Epub 2015/08/14. Eng.**

**2.Misdraji J. Mucinous epithelial neoplasms of the appendix and pseudomyxoma peritonei. Mod Pathol. 2015 Jan;28 Suppl 1:S67-79. PubMed PMID: 25560600.**

**3.Noguchi R, Yano H, Gohda Y, Suda R, Igari T, Ohta Y, et al. Molecular profiles of high-grade and low-grade pseudomyxoma peritonei. Cancer medicine. 2015 Dec;4(12):1809-16. PubMed PMID: 26475379. Pubmed Central PMCID: 5123786.**

**4.Nummela P, Saarinen L, Thiel A, Jarvinen P, Lehtonen R, Lepisto A, et al. Genomic profile of pseudomyxoma peritonei analyzed using next-generation sequencing and immunohistochemistry. Int J Cancer. 2015 Mar 01;136(5):E282-9. PubMed PMID: 25274248.**

**5.Singhi AD, Davison JM, Choudry HA, Pingpank JF, Ahrendt SA, Holtzman MP, et al. GNAS is frequently mutated in both low-grade and high-grade disseminated appendiceal mucinous neoplasms but does not affect survival. Human pathology. 2014 Aug;45(8):1737-43. PubMed PMID: 24925222.**

**6.Sio TT, Mansfield AS, Grotz TE, Graham RP, Molina JR, Que FG, et al. Concurrent MCL1 and JUN amplification in pseudomyxoma peritonei: a comprehensive genetic profiling and survival analysis. Journal of human genetics. 2014 Mar;59(3):124-8. PubMed PMID: 24369359. Pubmed Central PMCID: 3973125.**

**7.Liu X, Mody K, de Abreu FB, Pipas JM, Peterson JD, Gallagher TL, et al. Molecular profiling of appendiceal epithelial tumors using massively parallel sequencing to identify somatic mutations. Clinical chemistry. 2014 Jul;60(7):1004-11. PubMed PMID: 24821835.**

**8.Alakus H, Babicky ML, Ghosh P, Yost S, Jepsen K, Dai Y, et al. Genome-wide mutational landscape of mucinous carcinomatosis peritonei of appendiceal origin. Genome medicine. 2014;6(5):43. PubMed PMID: 24944587. Pubmed Central PMCID: 4062050.**

**9.Shih IM, Yan H, Speyrer D, Shmookler BM, Sugarbaker PH, Ronnett BM. Molecular genetic analysis of appendiceal mucinous adenomas in identical twins, including one with pseudomyxoma peritonei. The American journal of surgical pathology. 2001 Aug;25(8):1095-9. PubMed PMID: 11474297.**

**10.Racek AR, Rabe KG, Wick MJ, Psychogios A, Lindor NM. Primary appendiceal mucinous adenocarcinoma in two first-degree relatives: case report and review. Hereditary cancer in clinical practice. 2011 May 04;9(1):1. PubMed PMID: 21542938. Pubmed Central PMCID: 3098790.**

**11.*FastQC:* http://www.bioinformatics.babraham.ac.uk/projects/fastqc/**

**12.Martin M. Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnetjournal; Vol 17, No 1: Next Generation Sequencing Data Analysis. 2011 05/02/.**

**13.Li H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. 2013 (arXiv:1303.3997v2 [q-bio.GN].)**

**14.Picard: http://broadinstitute.github.io/picard**

**15.McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, et al. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing**

**data. Genome research. 2010 Sep;20(9):1297-303. PubMed PMID: 20644199. Pubmed Central PMCID: 2928508.**

**16.McLaren W, Pritchard B, Rios D, Chen Y, Flicek P, Cunningham F. Deriving the consequences of genomic variants with the Ensembl API and SNP Effect Predictor. Bioinformatics. 2010 Aug 15;26(16):2069-70. PubMed PMID: 20562413. Pubmed Central PMCID: 2916720.**

**17.Rios D, McLaren WM, Chen Y, Birney E, Stabenau A, Flicek P, et al. A database and API for variation, dense genotyping and resequencing data. BMC bioinformatics. 2010 May 11;11:238. PubMed PMID: 20459810. Pubmed Central PMCID: 2882931.**

**18.Genomes Project C, Abecasis GR, Auton A, Brooks LD, DePristo MA, Durbin RM, et al. An integrated map of genetic variation from 1,092 human genomes. Nature. 2012 Nov 01;491(7422):56-65. PubMed PMID: 23128226. Pubmed Central PMCID: 3498066.**

**19.Exome Variant Server. NHLBI GO Exome Sequencing Project (ESP), Seattle, WA (URL: http://evs.gs.washington.edu/EVS/) n.d.**

**20.Thompson ER, Doyle MA, Ryland GL, Rowley SM, Choong DYH, Tothill RW, et al. Exome Sequencing Identifies Rare Deleterious Mutations in DNA Repair Genes FANCC and BLM as Potential Breast Cancer Susceptibility Alleles. PLoS Genetics. 2012;8(9).**

**21.Ramensky V, Bork P, Sunyaev S. Human non-synonymous SNPs: server and survey. Nucleic acids research. 2002 Sep 01;30(17):3894-900. PubMed PMID: 12202775. Pubmed Central PMCID: 137415.**

**22.Ng PC, Henikoff S. SIFT: Predicting amino acid changes that affect protein function. Nucleic acids research. 2003 Jul 01;31(13):3812-4. PubMed PMID: 12824425. Pubmed Central PMCID: 168916.**

**23.Kircher M, Witten DM, Jain P, O'Roak BJ, Cooper GM, Shendure J. A general framework for estimating the relative pathogenicity of human genetic variants. Nat Genet. 2014 Mar;46(3):310-5. PubMed PMID: 24487276. Pubmed Central PMCID: 3992975.**

**24.Robinson JT, Thorvaldsdottir H, Winckler W, Guttman M, Lander ES, Getz G, et al. Integrative genomics viewer. Nat Biotech. 2011 01//print;29(1):24-6.**

**25.Thorvaldsdóttir H, Robinson JT, Mesirov JP. Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. Briefings in Bioinformatics. 2013;14(2):178-92.**

**26.Lek M, Karczewski KJ, Minikel EV, Samocha KE, Banks E, Fennell T, et al. Analysis of protein-coding genetic variation in 60,706 humans. Nature. 2016 08/18/print;536(7616):285-91.**

**27.Untergasser A, Cutcutache I, Koressaar T, Ye J, Faircloth BC, Remm M, et al. Primer3--new capabilities and interfaces. Nucleic acids research. 2012 Aug;40(15):e115. PubMed PMID: 22730293. Pubmed Central PMCID: 3424584.**

**28.Koressaar T, Remm M. Enhancements and modifications of primer design program Primer3. Bioinformatics. 2007;23(10):1289-91.**

**29.Thomas NA, Neville PJ, Baxter SW, Campbell IG. Genetic analysis of benign ovarian tumors. Int J Cancer. 2003 Jul 01;105(4):499-505. PubMed PMID: 12712441.**

**30.Kempers MJ, Kuiper RP, Ockeloen CW, Chappuis PO, Hutter P, Rahner N, et al. Risk of colorectal and endometrial cancers in EPCAM deletion-positive Lynch syndrome: a cohort study. The Lancet Oncology. 2011 Jan;12(1):49-55. PubMed PMID: 21145788. Pubmed Central PMCID: 3670774.**

**31.Palles C, Cazier JB, Howarth KM, Domingo E, Jones AM, Broderick P, et al. Germline mutations affecting the proofreading domains of POLE and POLD1 predispose to colorectal adenomas and carcinomas. Nat Genet. 2013 Feb;45(2):136-44. PubMed PMID: 23263490. Pubmed Central PMCID: 3785128.**

**32.Weren RD, Ligtenberg MJ, Kets CM, de Voer RM, Verwiel ET, Spruijt L, et al. A germline homozygous mutation in the base-excision repair gene NTHL1 causes adenomatous polyposis and colorectal cancer. Nat Genet. 2015 Jun;47(6):668-71. PubMed PMID: 25938944.**

**33.KOOREY D, BASHA NJ, TOMARAS C, FREIMAN J, ROBSON L, SMITH A. Appendiceal carcinoma complicating adenomatous polyposis in a young woman with a de novo constitutional reciprocal translocation t(5;8)(q22;p23.1). Journal of Medical Genetics. 2000;37(1):71-5.**

**34.Shin SM, Chung YJ, Oh ST, Jeon HM, Hwang LJ, Namkoong H, et al. HCCR-1-interacting molecule "deleted in polyposis 1" plays a tumor-suppressor role in colon carcinogenesis. Gastroenterology. 2006 Jun;130(7):2074-86. PubMed PMID: 16762630.**

**35.Ko J, Lee YH, Hwang SY, Lee YS, Shin SM, Hwang JH, et al. Identification and differential expression of novel human cervical cancer oncogene HCCR-2 in human cancers and its involvement in p53 stabilization. Oncogene. 2003 Jul 24;22(30):4679-89. PubMed PMID: 12879013.**

**36.Burke B. PREEParing for mitosis. Developmental cell. 2013 Aug 12;26(3):221-2. PubMed PMID: 23948250.**

**37.Camargo AC, Gomes MD, Reichl AP, Ferro ES, Jacchieri S, Hirata IY, et al. Structural features that make oligopeptides susceptible substrates for hydrolysis by recombinant thimet oligopeptidase. The Biochemical journal. 1997 Jun 01;324 ( Pt 2):517-22. PubMed PMID: 9182712. Pubmed Central PMCID: PMC1218460. Epub 1997/06/01. eng.**

**38.Nomoto S, Hishida M, Inokawa Y, Takano N, Kanda M, Nishikawa Y, et al. Expression analysis of THOP1 in background liver, a prognostic predictive factor in hepatocellular carcinoma, extracted by multiarray analysis. Ann Surg Oncol. 2014 Jun;21 Suppl 3:S443-50. PubMed PMID: 24604581. Epub 2014/03/08. eng.**

**39.Cheng T-L, Lai C-H, Jiang S-J, Hung J-H, Liu S-K, Chang B-I, et al. RHBDL2 Is a Critical Membrane Protease for Anoikis Resistance in Human Malignant Epithelial Cells. The Scientific World Journal. 2014;2014:8.**

**40.Katoh M, Nakagama H. FGF Receptors: Cancer Biology and Therapeutics. Medicinal Research Reviews. 2014;34(2):280-300.**

**41.Xu W, Li Y, Wang X, Chen B, Wang Y, Liu S, et al. FGFR4 transmembrane domain polymorphism and cancer risk: A meta-analysis including 8555 subjects. European journal of cancer. 2010 2010/12/01/;46(18):3332-8.**

**42.Lynch TJ, Bell DW, Sordella R, Gurubhagavatula S, Okimoto RA, Brannigan BW, et al. Activating mutations in the epidermal growth factor receptor underlying responsiveness of non-small-cell lung cancer to gefitinib. The New England journal of medicine. 2004 May 20;350(21):2129-39. PubMed PMID: 15118073. Epub 2004/05/01. eng.**

**43.Rosell R, Carcereny E, Gervais R, Vergnenegre A, Massuti B, Felip E, et al. Erlotinib versus standard chemotherapy as first-line treatment for European patients with advanced EGFR mutation-positive non-small-cell lung cancer (EURTAC): a multicentre, open-label, randomised phase 3 trial. The Lancet Oncology. 2012 Mar;13(3):239-46. PubMed PMID: 22285168. Epub 2012/01/31. eng.**

**44.Zhou C, Wu YL, Chen G, Feng J, Liu XQ, Wang C, et al. Erlotinib versus chemotherapy as first-line treatment for patients with advanced EGFR mutation-positive non-small-cell lung cancer (OPTIMAL, CTONG-0802): a multicentre, open-label, randomised, phase 3 study. The Lancet Oncology. 2011 Aug;12(8):735-42. PubMed PMID: 21783417. Epub 2011/07/26. eng.**

**45.Mok TS, Wu Y-L, Thongprasert S, Yang C-H, Chu D-T, Saijo N, et al. Gefitinib or Carboplatin–Paclitaxel in Pulmonary Adenocarcinoma. New England Journal of Medicine. 2009;361(10):947-57. PubMed PMID: 19692680.**

### *4.3.5.2. Candidate Gene Approach*

#### 4.3.5.2.1. Genes in known CRC pathways

Twenty-three LoF variants in 22 CRC pathway genes were identified from the WES data from the discovery cohort. Two variants with a QUAL score <100 and a variant frequency of <0.3 were omitted. For genes with a LoF variant within the DNA repair pathway, an in-house case control dataset that utilised targeted gene sequencing was used to further evaluate these genes. A dataset of 1061 Australian breast cancer cases and 1059 controls (Li et al., 2016) was screened for variants in DNA repair genes using the Haloplex Target Enrichment Kit. This dataset was used to omit variants in DNA repair genes with an incidence of >0.5% in local controls. A variant in *EXO1* that was observed in 7 controls, and a variant in *CEP164* that was observed in 6 controls, were omitted. These 2 variants were not observed in ClinVar. All remaining LoF variants were visualised in the Integrated Genome Viewer (Robinson et al., 2011). One variant from a low quality site was omitted. An *MSH6* LoF variant (ENST00000234420.4:c.4001+2_4001+5delTAAC) that did not affect the essential splice site (discussed in Chapter 4.3.1.3) was omitted.

Sanger sequencing was used to validate the remaining 17 LoF variants in pathway genes. The variant prioritisation process is summarised in Figure 4-7.Fifteen variants were detected on Sanger sequencing (Table 4.16). The minimum estimate of LoF frequency in each of these genes in the ExAC-TCGA dataset is shown.

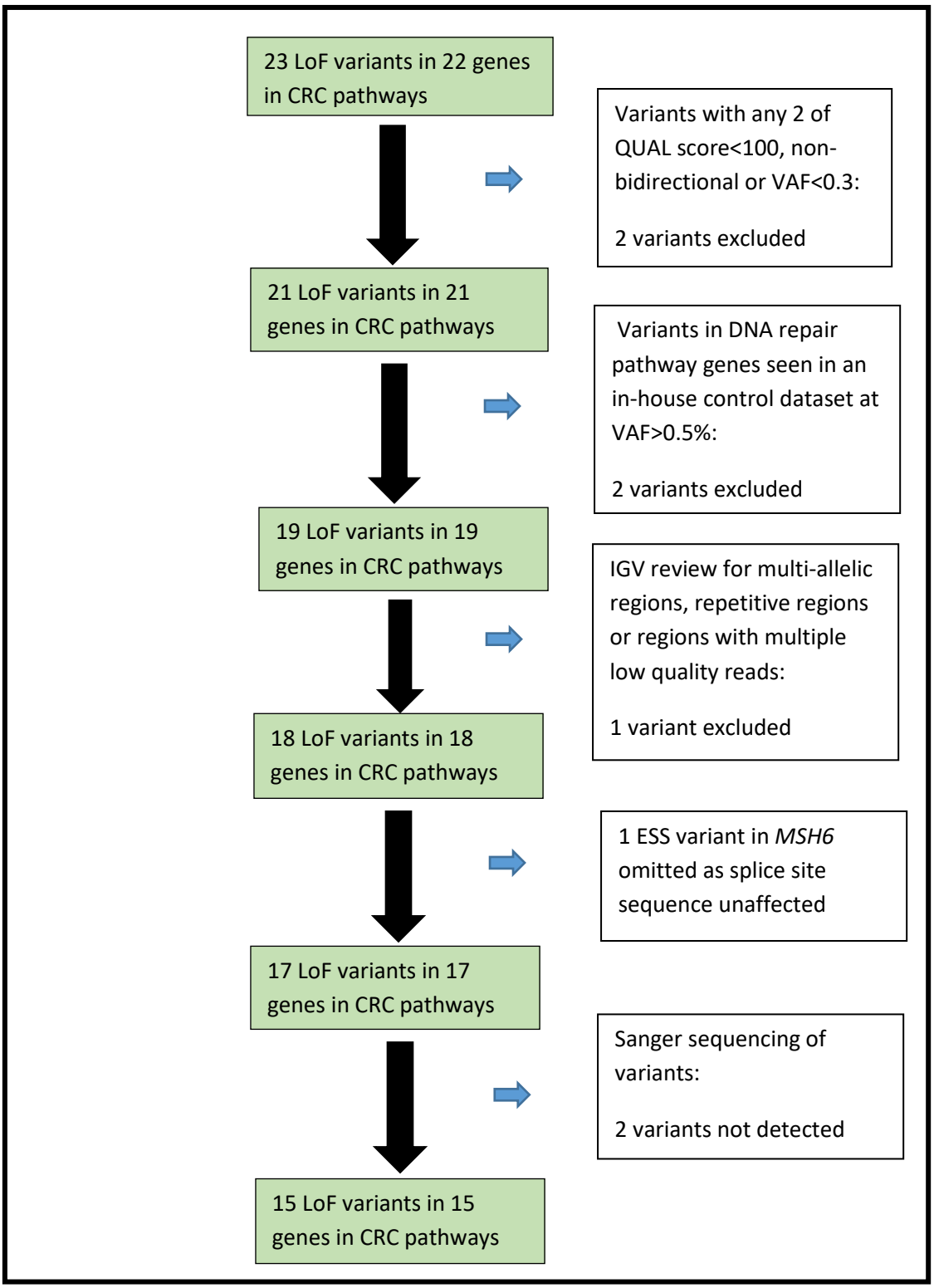



**Figure 4-7. Variant filtering pipeline for candidate CRC-predisposing genes in CRC pathways.**

LoF, loss-of-function. CRC, colorectal cancer. VAF, variant allele frequency. IGV, Integrated Genome Viewer. ESS, essential splice site.

**Table 4-17. LoF variants in 15 CRC pathway genes validated by Sanger sequencing**

| Gene | Gene Description | CRC pathways | Ensembl Gene | Genomic position | cDNA position | Ensembl protein | Amino acid change | ExAC-TCGA AF | #Sample | Minimum estimate of percentage of individuals with a LoF variant in Non-TCGA Subset of ExAC (%) |
|---|---|---|---|---|---|---|---|---|---|---|
| ***WNT10A*** | wingless-type MMTV integration site family, member 10A | Wnt | ENSG00000135925 | 2:219747090C>A | c.321C>A | ENSP00000258411.3 | p.Cys107Ter | 5.27E-04 | PC18 | 0.13 |
| ***LRRFIP2*** | leucine rich repeat (in FLII) interacting protein 2 | Wnt | ENSG00000093167 | 3:37154433delC | c.411delG | ENSP00000392217.1 | p.Arg137SerfsTer10 | . | PC54 | 0.13 |
| ***NOTCH4*** | notch 4 | Notch | ENSG00000204301 | 6:32165201G>A | c.4927C>T | ENSP00000364163.3 | p.Arg1643Ter | 1.60E-04 | PC41 | 1.19 |
| ***EP300*** | E1A binding protein p300 | Notch | ENSG00000100393 | 22:41523489_41523490delAG | c.907-2_907-1delAG | . | . | . | PC16 | 0.02 |
| ***CNTN6*** | contactin 6 | Notch | ENSG00000134115 | 3:1424850delT | c.2391delT | ENSP00000341882.2 | p.Glu799LysfsTer53 | . | PC12 | 0.30 |
| ***JAG2*** | jagged 2 | Notch | ENSG00000184916 | 14:105609320_105609321insC | c.3314_3315insG | ENSP00000328566.2 | p.His1106ProfsTer35 | . | PC36 | 0.01 |
| ***XRCC6BP1*** | XRCC6 binding protein 1 | DNA repair | ENSG00000166896 | 12:58350658T>G | c.726T>G | ENSP00000300145.3 | p.Tyr242Ter | . | PC38 | 0.04 |
| ***APLF*** | aprataxin and PNKP like factor | DNA repair | ENSG00000169621 | 2:68694913delC | c.50delC | ENSP00000307004.4 | p.Leu18TrpfsTer7 | 3.89E-05 | PC2, PC3 | 2.10 |
| ***DMC1*** | DNA meiotic recombinase 1 | DNA repair | ENSG00000100206 | 22:38935398delC | c.514delG | ENSP00000216024.2 | p.Asp172ThrfsTer8 | . | PC44 | 0.02 |
| ***MUS81*** | MUS81 structure-specific endonuclease subunit | DNA repair | ENSG00000172732 | 11:65628316_65628329delAGGCGACCCGCAGC | c.92_105delAGGCGACCCGCAGC | ENSP00000307853.4 | p.Thr33ProfsTer27 | . | PC29 | 0.08 |

**Table 4-17 (continued). LoF variants in 15 CRC pathway genes validated by Sanger sequencing**

| Gene | Gene Description | CRC pathways | Ensembl Gene | Genomic position | cDNA position | Ensembl protein | Amino acid change | ExAC-TCGA AF | #Sample | Minimum estimate of percentage of individuals with a LoF variant in Non-TCGA Subset of ExAC (%) |
|---|---|---|---|---|---|---|---|---|---|---|
| ***USP28*** | ubiquitin specific peptidase 28 | DNA repair | ENSG00000048028 | 11:113724999_113725003delAAGG | c.113_116delTTCC | ENSP00000003302.4 | p.Ser38PhefsTer6 | . | PC18 | 0.05 |
| ***ERCC3*** | excision repair cross-complementing rodent repair deficiency, complementation group 3 | DNA repair | ENSG00000163161 | 2:128030505_128030506insC | c.1762_1763insG | ENSP00000285398.2 | p.Glu588GlyfsTer16 | 4.71E-05 | PC34 | 0.22 |
| ***POLN*** | polymerase (DNA directed) nu | DNA repair | ENSG00000130997 | 4:2178059delA | c.1182delT | ENSP00000372316.1 | p.Gln395ArgfsTer7 | . | PC27 | 1.83 |
| ***BRCA2*** | Breast cancer 2, early onset | DNA repair | ENSG00000139618 | 13:32906916_32906919delAAAG | c.1301_1304delAAAG | ENSP00000439902.1 | p.Lys437IlefsTer22 | | PC10 | 1.67 |
| ***GDF2*** | growth differentiation factor 2 | BMP | ENSG00000128802 | 10:48414131_48414132delGG | c.736_737delCC | ENSP00000249598.1 | p.Pro246ArgfsTer14 | . | PC26 | 0.01 |

### *4.3.5.3. Highly constrained Genes*

Thirteen known CRC-predisposing genes were annotated with ExAC pLI (Table 4.17), missense z-scores (Table 4.17) and RVIS percentiles (Table 4.18) in order to assess their utility in identifying CRC-predisposing genes.

**Table 4-18. ExAC pLI scores of known CRC-predisposing genes. A pLI of ≥0.9 indicates a LoF-intolerant gene.**

| Gene | ExAC pLI |
|---|---|
| *APC* | 1.00 |
| *STK11* | 0.98 |
| *SMAD4* | 0.97 |
| *MSH2* | 0.87 |
| *BMPR1A* | 0.77 |
| *MLH1* | 0.74 |
| *GREM1* | 0.71 |
| *MSH6* | pLI not provided |
| *PMS2* | 0.00 |
| *POLD1* | 0.00 |
| *POLE* | 0.00 |
| *EPCAM* | 0.00 |
| *MUTYH* | 0.00 |

pLI, probability of being loss-of-function intolerant.

Three genes showed an ExAC pLI>0.9. *PMS2* did not show constraint using the pLI, which may be due to the presence of pseudogenes increasing the number of LoF variants called. Mutations in *MUTYH* predispose to CRC in an autosomal recessive state and hence there was no constraint using pLI. Only 3'deletions of the *EPCAM* gene have been reported to predispose to CRC, through transcriptional read-through into the neighbouring *MSH2* gene, resulting in methylation of the latter (Ligtenberg et al., 2009). This is consistent with *EPCAM* being tolerant of LoF variants at a gene-wide level.

**Table 4-19. ExAC missense z scores of known CRC-predisposing genes. A missense z-score of >3.09 indicates a significant constraint on missense variants.**

| Gene | Missense z score |
|---|---|
| *SMAD4* | 4.18 |
| *BMPR1A* | 2.81 |
| *POLD1* | 2.79 |
| *GREM1* | 1.94 |
| *STK11* | 1.71 |
| *POLE* | 1.57 |
| *PMS2* | 0.04 |
| *MUTYH* | -0.46 |
| *MLH1* | -1.09 |
| *APC* | -1.49 |
| *EPCAM* | -1.93 |
| *MSH2* | -4.15 |
| *MSH6* | z-score not provided |

*POLE* and *POLD1* mutations that predispose to CRC are usually missense mutations within the exonuclease region. While the ExAC missense z-scores were positive values for both genes, the missense z-score did not exceed 3.09 (Table 4.18). *SMAD4* showed significant constraint of missense variants with a missense z-score exceeding 3.09. As the use of missense z-scores with a cut-off of >3.09 did not identify known CRC-predisposing genes with predominantly missense mutations, missense z-scores were not used for identification of novel CRC-predisposing genes.

**Table 4-20. RVIS percentiles of the known CRC-predisposing genes, with the lowest percentiles indicating genes intolerant of common functional variation**

| Gene | RVIS percentile |
|---|---|
| *APC* | 0.9 |
| *MSH6* | 1.28 |
| *MSH2* | 1.68 |
| *POLD1* | 11.54 |
| *POLE* | 12.09 |
| *STK11* | 20.54 |
| *BMPR1A* | 26.53 |
| *SMAD4* | 31.69 |
| *MUTYH* | 62.1 |
| *GREM1* | 62.74 |
| *EPCAM* | 75.29 |
| *MLH1* | 80.3 |
| *PMS2* | 95.29 |

RVIS, Residual Variation Intolerance Score. RVIS percentiles given to two decimal places.

With regards to RVIS percentiles, the known CRC-predisposing genes showed varying tolerance to common functional variation. APC was among the 1% most intolerant of genes, while *MLH1* and *PMS2* were tolerant at 80 and 95% RVIS percentiles respectively (Table 4.19). While the reference paper suggested using known disease-predisposing genes to determine the appropriate RVIS percentile cut-off, in this case there was no appropriate cut-off due to the spread of the known CRC-predisposing genes across RVIS percentiles.

An ExAC pLI of ≥0.9 identified 59 genes with one or more LoF variants in the discovery dataset that were intolerant of LoF variants. Of these 59 genes, a LoF variant was seen only once in the discovery dataset in 52 genes. Twelve of these genes were omitted, as these were from PC47 that had an excess of LoF variants as described in Section 4.3.3.

Seven genes had recurrent LoF variants in constrained genes. Six of these genes had been omitted following analysis of the variants in Section 4.3.5.1.1 for the following reasons: a combination of low variant QUAL score/variant frequency or non-bidirectional reads, the variants were long deletions, on low quality reads or because the variant occurred too frequently in the discovery dataset. Manual inspection of each of these variants in the ExAC browser subsequently found that all variants except one had a VQSRTrancheINDEL of 96.00 or higher. The variant quality score recalibration (VQSR) method is a quality score that uses machine learning on highly validated variants to evaluate multiple variant parameters in order to differentiate true variants from false positives. VQSRTrancheINDEL scores of 95.00 or higher corresponded to tranches with more false positives. One variant was not observed in the ExAC dataset.

One of the genes with a recurrent LoF variant was seen in 2 affected individuals from the same family. This variant was omitted during the intra-family analysis (Section 4.3.5.1.2.2) as it had

an allele frequency of >0.1% in the Non-TCGA Subset of the ExAC dataset. This variant had a VQSRTrancheINDEL of 99.95to100.00. None of the genes with recurrent LoF variants in this analysis were taken forward to validation.

Of the 40 ExAC constrained genes with a LoF variant in one sample, 10 had 2 of the following: QUAL score<100, variant allele frequency of <0.3 or were not seen on reads bi-directionally. Thirty ExAC constrained genes with a LoF variant in one individual from the discovery dataset remained (Supplementary Table A-2).

Among these genes were *EP300* and *JAG2* that had been identified in the pathway analysis. *EP300* had a pLI of 1, while *JAG2* had a pLI of 0.99. A literature search for each of the 30 genes was performed to identify any known functions of each gene. Due to constraints on the number of genes that could be taken forward for validation using a targeted sequencing panel (Chapter 5), and because these genes had a LoF variant in only one individual in the ExAC dataset, only 3 ExAC constrained genes (*INCENP, PHRF1 and SIN3B)* were selected for validation due to plausible CRC predisposition.

INCENP forms part of the chromosomal passenger complex that regulates chromosome segregation and cytokinesis during mitosis (Carmena et al., 2012). INCENP is important for the localisation of the spindle assembly checkpoint protein, Bub1, to the kinetochore (Qi et al., 2006). Importantly, germline *BUB1* mutations have been found in young-onset CRC (de Voer et al., 2013). PHRF1 is an ubiquitin ligase that may function as a tumour suppressor gene through its interaction with TGFβ signalling (Ettahar et al., 2013). SIN3B is a transcription regulator that down-regulates Myc expression (Garcia-Sanz et al., 2014).

**Table 4-21.Three constrained genes with a LoF variant, selected due to plausible CRC predisposition**

| Gene | Gene Description | Ensembl Gene | Genomic position | Ensembl Transcript | cDNA position | Ensembl protein | Amino acid change | ExAC pLI | ExAC-TCGA AF | #Sample | Minimum estimate of percentage of individuals with a LoF variant in Non-TCGA Subset of ExAC |
|---|---|---|---|---|---|---|---|---|---|---|---|
| ***INCENP*** | inner centromere protein antigens 135/155kDa | ENSG00000149503 | 11:61919405delC | ENST00000278849 | c.2702delC | ENSP00000278849.4 | p.Arg902GlyfsTer10 | 0.99 | . | PC18 | 0.03 |
| ***PHRF1*** | PHD and ring finger domains 1 | ENSG00000070047 | 11:608400_608401delCA | ENST00000264555 | c.2944_2945delCA | ENSP00000264555.5 | p.His982GlnfsTer42 | 0.94 | . | PC49 | 0.35 |
| ***SIN3B*** | SIN3 transcription regulator homolog B (yeast) | ENSG00000127511 | 19:16962280_16962281 insC | ENST00000379803 | c.784_785insC | ENSP00000369131.1 | p.Glu264GlyfsTer49 | 0.99 | . | PC35 | 0.11 |

## 4.4. Discussion

Using a customised filtering pipeline to first identify PPVs, gene level (inter-family and gene constraint), variant level (intra-family) and pathway level analyses were then performed to prioritise 46 candidate CRC-predisposing genes (Table 4.21). The final candidate gene list comprised of 12 genes from the inter-family analysis, 16 genes from the intra-family analysis, 15 genes from the CRC pathway analysis and 3 genes from the gene constraint analysis. LoF variants were used for candidate CRC-predisposing gene identification, apart from in the intra-family analysis that included shared private missense variants predicted to be deleterious using CADD. Despite the use of CADD to prioritise potentially pathogenic missense variants, the number of missense variants per exome was approximately one order of magnitude higher than the number of LoF variants per exome.

**Table 4-22. Forty-six candidate CRC-predisposing genes**

| Gene | Reason for inclusion | cDNA | amino acid change | Gene Function Summary (based on GeneCards® Human Gene Database (Crown Human Genome Center, 1997)) |
|---|---|---|---|---|
| ***SERPINB7*** | inter-family analysis (different variants) | c.639_640delGT; c.862A>T | p.Lys213AsnfsTer7; p.Lys288Ter | protease inhibitor |
| ***GALC*** | inter-family analysis (different variants) | c.84_85insCCGCG; c.40_41insC | p.Val29ProfsTer45; p.Leu14ProfsTer14 | catabolises galactosylceramide, a major lipid in myelin, kidney and epithelial cells of small and large intestine |
| ***GALNT8*** | inter-family analysis (different variants) | c.1357C>T; c.1360-1G>A | p.Gln453Ter; - | involved in O-linked glycosylation in Golgi apparatus |
| ***GPR144*** | inter-family analysis (different variants) | c.959G>A; c.2836_2837delCT | p.Trp320Ter; p.Phe947SerfsTer4 | unknown |
| ***PDE5A*** | inter-family analysis (different variants) | c.2253_2256delGAGA; c.19A>T | p.Arg752LysfsTer5; p.Lys7Ter | formation of 5'-GMP from cGMP by hydrolysis |
| ***MS4A15*** | inter-family analysis (different variants) | c.82delG; c.333+1G>C | p.Gly28AspfsTer16; - | possible role in signal transduction |
| ***ZNF788*** | inter-family analysis (different variants) | c.765delG; c.1137_1138delAT | p.Glu256AsnfsTer13; p.Cys380Ter | possible role in regulation of transcription |
| ***RIPPLY3*** | inter-family analysis (same variant) | c.292C>T | p.Gln98Ter | Negative regulation of *TBX1* (that is involved in development of the pharynx) |
| ***ASCL3*** | inter-family analysis (same variant) | c.382C>T | p.Arg128Ter | involved in tissue development and differentiation |
| ***TEX14*** | inter-family analysis (same variant) | c.1003C>T | p.Arg335Ter | essential for kinetochore-microtubule attachment in mitosis |
| ***SH2D4A*** | inter-family analysis (same variant) | c.706+1G>A | - | inhibits oestrogen-induced cell proliferation |
| ***TNFRSF10B*** | inter-family analysis (same variant) | c.748+1G>A | - | part of an apoptosis signalling complex |
| **NEDD4** | intra-family analysis | c.238-1G>A | - | ubiquitin ligase |
| **TIMELESS** | intra-family analysis | c.3553+2_3553+5delGGAT | - | regulates DNA replication in stress and normal cell conditions |
| ***INPP5E*** | intra-family analysis | c.1438T>C | p.Phe480Leu | regulates intracellular calcium |
| ***GOLGA5*** | intra-family analysis | c.2041G>A | p.Asp681Asn | Golgi apparatus protein. Translocations involving *RET* have been found in tumours |

**Table 4-21 (continued). Forty-six candidate CRC-predisposing genes**

| Gene | Reason for inclusion | cDNA | amino acid change | Gene Function Summary (based on GeneCards® Human Gene Database (Crown Human Genome Center, 1997)) |
|---|---|---|---|---|
| ***PLSCR3*** | intra-family analysis | c.776C>T | p.Pro259Leu | may be involved in fibrin clot formation, mast cell activation, and apoptosis |
| ***HTR5A*** | intra-family analysis | c.361G>A | p.Asp121Asn | serotonin receptor |
| ***HCN3*** | intra-family analysis | c.932G>C | p.Cys311Ser | potassium channel |
| ***GPR137B*** | intra-family analysis | c.448G>C | p.Glu150Gln | unknown |
| ***SF3A3*** | intra-family analysis | c.1409G>A | p.Arg470Gln | subunit of splicing factor SF3A |
| ***FKBP10*** | intra-family analysis | c.1256C>T | p.Ser419Leu | molecular chaperone involved in protein folding |
| ***IL4R*** | intra-family analysis | c.2164G>A | p.Gly722Ser | IL4/IL13 receptor |
| ***LRRK2*** | intra-family analysis | c.2414T>A | p.Leu805His | autophagy regulation |
| ***SUCLG2*** | intra-family analysis | c.1045C>T | p.Leu349Phe | binds succinate in the citric acid cycle |
| ***UFL1*** | intra-family analysis | c.944A>C | p.Glu315Ala | involved in protein post-translational modification |
| ***MMP2*** | intra-family analysis | c.1591A>G | p.Lys531Glu | zinc-dependent endopeptidase involved in extracellular matrix degradation |
| ***TMCC3*** | intra-family analysis | c.362A>G | p.Lys121Arg | unknown |
| ***WNT10A*** | Gene in CRC pathway (Wnt) | c.321C>A | p.Cys107Ter | ligand for transmembrane receptors of the frizzled family |
| ***LRRFIP2*** | Gene in CRC pathway (Wnt) | c.411delG | p.Arg137SerfsTer10 | may activate canonical Wnt pathway |
| ***NOTCH4*** | Gene in CRC pathway (Notch) | c.4927C>T | p.Arg1643Ter | receptor involved in cell fate determination |
| ***EP300*** | Gene in CRC pathway (Notch); high gene constraint score | c.907-2_907-1delAG | . | histone acetyltransferase with core functions in cell proliferation and differentiation |
| ***CNTN6*** | Gene in CRC pathway (Notch) | c.2391delT | p.Glu799LysfsTer53 | NOTCH1 ligand |
| ***JAG2*** | Gene in CRC pathway (Notch); high gene constraint score | c.3314_3315insG | p.His1106ProfsTer35 | Notch ligand |
| ***XRCC6BP1*** | Gene in CRC pathway (DNA repair) | c.726T>G | p.Tyr242Ter | increases double-strand break repair efficiency |
| ***APLF*** | Gene in CRC pathway (DNA repair) | c.50delC | p.Leu18TrpfsTer7 | nuclease involved in single strand and double strand DNA break repair |
| ***DMC1*** | Gene in CRC pathway (DNA repair) | c.514delG | p.Asp172ThrfsTer8 | involved in meiotic homologous recombination |

**Table 4-21 (continued). Forty-six candidate CRC-predisposing genes**

| Gene | Reason for inclusion | cDNA | amino acid change | Gene Function Summary (based on GeneCards® Human Gene Database (Crown Human Genome Center, 1997)) |
|---|---|---|---|---|
| ***MUS81*** | Gene in CRC pathway (DNA repair) | c.92_105delAGGCGACCCGCAGC | p.Thr33ProfsTer27 | endonuclease involved in DNA double-strand break repair |
| ***USP28*** | Gene in CRC pathway (DNA repair) | c.113_116delTTCC | p.Ser38PhefsTer6 | Deubiquitinase involved in DNA damage response/apoptosis |
| ***ERCC3*** | Gene in CRC pathway (DNA repair) | c.1762_1763insG | p.Glu588GlyfsTer16 | 3'-5' DNA helicase involved in nucleotide excision repair |
| ***POLN*** | Gene in CRC pathway (DNA repair) | c.1182delT | p.Gln395ArgfsTer7 | Error-prone DNA polymerase |
| ***BRCA2*** | Gene in CRC pathway (DNA repair) | c.1301_1304delAAAG | p.Lys437IlefsTer22 | homologous recombination and double-strand DNA break repair |
| ***GDF2*** | Gene in CRC pathway (BMP) | c.736_737delCC | p.Pro246ArgfsTer14 | TGFβ ligand |
| ***INCENP*** | High gene constraint score | c.2702delC | p.Arg902GlyfsTer10 | involved in mitosis regulation |
| ***PHRF1*** | High gene constraint score | c.2944_2945delCA | p.His982GlnfsTer42 | unknown |
| ***SIN3B*** | High gene constraint score | c.784_785insC | p.Glu264GlyfsTer49 | transcriptional repressor that antagonises *MYC* |

### 4.4.1. Implications of identifying pathogenic variants in known CRC-predisposing genes through WES

The identification of 3 inherited pathogenic variants in known CRC-predisposing genes in 3 individuals from this cohort validates the selection criteria and analysis pipeline used to establish the discovery cohort. Furthermore, the identification of pathogenic variants through germline WES, when no predisposing mutation had been identified in the FCC has potential clinical implications.

The *PMS2* mutation identified in PC15 is a common *PMS2* mutation predisposing to Lynch syndrome (Clendenning et al., 2008). Lack of a strong family history is consistent with the lower penetrance of a *PMS2* mutation. MMR IHC for PMS2 in this individual did not show an absence of staining. It is unclear why this was the case, although it is possible that the IHC result is a false positive. Microsatellite testing for 4 mononucleotide repeat markers (CAT25, bat25, bat26, bat 40) showed stability at all 4 markers. Microsatellite stable CRC (based on assessment of 5 mononucleotide markers) has been seen with germline *MSH6* mutations (Giráldez et al., 2010). In a recent population-based panel testing study in young-onset familial CRC that included the mismatch repair genes, the authors commented that the cost of MMR IHC plus Sanger sequencing is comparable to that of exome sequencing if at least 3 genes are tested (Chubb et al., 2015), although the cost of interpretation of exome sequencing data may not have been factored in. Nonetheless, as the cost of next generation sequencing continues to fall, panel testing may be the way forward.

PC50, who was found to have a *BMPR1A* mutation, presented to an Australian FCC with rectal cancer at the age of 48 years reporting a history of a total colectomy overseas at the age of 14 years for colonic polyposis. Genetic testing for *APC* and *MUTYH* mutations was uninformative. Differentials for childhood colonic polyposis may not have been considered as this is usually managed in paediatric centres, and the polyps had been reported by the patient as adenomas. Panel testing for genes that predispose to colonic polyposis may be helpful in situations where full histopathological information is unavailable, for example in this case when the initial surgical management was performed overseas.

PC51 was strongly suspected of having FAP despite a negative PTT for APC. This was a historic sample that had not undergone current accepted levels of diagnostic testing. A recurring review of individuals with a clinical phenotype and/or family history strongly suggestive of an inherited pathogenic variant in a CRC-predisposing gene can prove beneficial as new technologies for genetic testing emerge. This review would take into account better laboratory techniques in analysing known CRC genes (for example, pathogenic intronic mutations in *APC*) as well as emerging CRC-predisposing genes for particular clinical syndromes as they are discovered. While such a review system would require additional time and resources, the benefits of identifying a family mutation predisposing to CRC are significant and would allow for: 1)appropriate counselling on risk reduction strategies that have, in the case of familial adenomatous polyposis due to inherited pathogenic variants in *APC* and Lynch syndrome, been shown to increase overall survival for screened individuals (Heiskanen et al., 2000a, Renkonen-Sinisalo et al., 2000b), and 2) segregation of the mutation within the family and hence exclusion of family members who do not carry the family mutation from rigorous screening procedures (which may improve the use of public health resources).

### 4.4.2. Variants of uncertain significance in known CRC-predisposing genes

Two *APC* variants with concordant *in silico* predictions of deleteriousness in Condel and CADD were observed in individuals without known colonic polyposis. While CADD incorporates SIFT and PolyPhen, two in silico predictors used by Condel to generate a consensus score, it also relies on more than 60 other annotations. It is possible that these variants predispose to an increased risk of CRC in the absence of polyposis, akin to the *APC* p.I1307K variant, a founder mutation in the Jewish population (Shtoyerman-Chen et al., 2001) associated with a moderately increased risk of CRC (Boursi et al., 2013). Further supporting information including segregation and case-control studies is required for interpretation of these variants and since neither individual had affected family members it was assumed these variants were not pathogenic and these were included in the discovery cohort.

### 4.4.3. Identification of PPVs

The filtering pipeline for identifying PPVs used global population databases to identify rare alleles. A gene level analysis was then performed to exclude genes with an excess of LoF variants in a local population, as PPVs in these genes are less likely to affect function if they are commonly seen in a population. Finally, CADD was used to predict the pathogenicity of the rare missense variants and select for the top 10% most pathogenic variants.

Filtering variants using global population allele frequencies resulted in the identification of an individual with an excess of rare missense variants. On further analysis, these variants were common in the African population within the 1000G project, which may be due to the individual having African ancestry. In order to minimise this effect when the ethnicity of the individuals in a study is not known, a reasonable future approach would be to filter out variants using the highest allele frequency seen in any sub-population within the population databases. This is useful if the individual is from a population that is rare in the local population but represented as a sub-group in a global population database.

The addition of the institutional WES dataset was able to reduce the number of rare variants by one third after filtering using the global databases. As the Australian population is predominantly of European descent that is well represented in all the global databases, this may be due to removal of alignment artefacts rather than population alleles. The selection of the top 10% most deleterious missense variants using CADD still left more than two thirds of the rare missense alleles. Missense variants that are rare are mildly deleterious in up to 70% of cases (Kryukov et al., 2007), hence the CADD score may be most useful for ranking individual missense variants for further evaluation, as was done in the intra-family analysis. The CADD scaled scores for *POLE* p.Leu424Val (NM_006231; c.1270C>G) and *POLD1* p.Ser478Asn (NM_002691: c.1433G>A), the 2 autosomal dominant missense mutations initially identified in polymerase-proofreading associated polyposis families (Palles et al., 2013) are 15.39 and 25.2 respectively. Hence a CADD scaled score cut-off of 10 is reasonable as it would have kept these pathogenic variants in for further analysis.

In a discovery study such as this, data filters that are more inclusive at the beginning are favoured over stringent filters with high specificity. However, in the later parts of the analyses, more stringent filters such as a higher variant frequency cut-off and a lower population

frequency cut-off were employed to prioritise the variants most likely to be deleterious. This approach has also been used in another WES study of familial CRC (Esteban-Jurado et al., 2014).

### 4.4.4. Strengths and limitations of the analysis methods

Four analyses were performed on the identified PPVs from the discovery cohort: 1) inter-family analysis, 2) intra-family analysis, 3) candidate genes within CRC pathways, and 4) gene constraint.

The inter-family analysis provided an agnostic approach to candidate gene identification, but relied on a candidate gene being recurrently mutated in a cohort of 51 individuals affected with CRC. The intra-family analysis was limited by the close relatedness (FDRs and SDRs) of the affected individuals in each family resulting in a large number of shared variants, potentially by chance. The FCCTX family (PC41, PC54) had the potential to be more informative, with 2 affected third degree relatives. However, while the proband (PC41), his paternal uncle (PC54), father and non-affected twin sister agreed to participate in the study, his affected paternal aunt and paternal cousins did not respond to invitations to participate in this study. Interestingly, PC41 and PC54 had different clinical presentations, with PC54 presenting with screen-detected colonic adenomatous polyps at the age of 56 years while PC41 and his affected father presented with CRC at the ages of 38 and 61 years respectively without synchronous adenomatous polyps. Intra-familial phenotypic heterogeneity has been described for *APC* (Rozen et al., 1999a), hence it was postulated that a mutation in the same gene predisposed PC41 and PC54 to CRC and adenomatous polyps (pre-cancer) respectively. It is possible however that PC41 and PC54 have mutations in different novel CRC-predisposing genes.

The analysis of the father and daughter with AMTs identified 16 novel shared variants with no loss of the corresponding wildtype allele in tumour tissue. No inherited pathogenic variant in a known CRC-predisposing gene was identified, suggesting that familial AMTs have a different genetic aetiology to more common subtypes of familial CRC. It is possible that the AMTs arose in this family by chance, although the rarity of this tumour would go against this.

The candidate gene analysis identified genes within pathways implicated in CRC and since LoF variants in these pathway genes were each seen in only one family in the discovery cohort they would not have been identified through the inter-family analysis. A limitation of this approach is that the impact of perturbation of a particular gene on a pathway is not always known from the literature.

Gene constraint scores are an attractive tool for prioritising genes that are highly conserved and hence likely to cause disease when mutated. While a PPV in a constrained gene may have deleterious effects unrelated to cancer predisposition, constrained genes indicate that a LoF variant is under negative selection. The finding of 2 constrained genes (*EP300* and *JAG2*) identified through the CRC pathways analysis lent further support to these genes being candidate CRC-predisposing genes.

Genes with a LoF variant identified in a significant proportion of the population are unlikely to predispose to disease. This was examined at 2 stages during the analyses. The first occasion

was during the identification of PPVs, when a relatively high cut-off of 5% of a population of 198 in-house exomes was used. When the Non-TCGA Subset of ExAC became available, a minimum estimate of the percentage of the population within the Non-TCGA Subset of ExAC with a LoF variant in a gene could be calculated. Setting a cut-off for this estimate was complicated by genes in which the majority of LoF variants occurred in the last exon, which may not result in nonsense-mediated decay. Further inspection of the distribution of LoF variants in genes with a LoF variant in 2-4% of the Non-TCGA Subset of ExAC was required to ensure that genes with LoF variants predominantly in the last exon were not excluded from further analysis.

Synonymous variants and variants in intronic regions surrounding essential splice sites were not considered in any of the analyses, while missense variants were only utilised in the intra-family analysis. While it is possible that these variant types may be deleterious in some of the affected individuals, the difficulty in distinguishing functionally deleterious variants from benign variants in these variant classes limited their use in identifying candidate CRC-predisposing genes.

### 4.4.5. Functions of the strongest candidate CRC-predisposing genes

Three candidate CRC-predisposing genes were identified through 2 out of the 4 analysis methods employed (inter-family analysis, intra-family analysis, CRC pathway analysis, gene constraint analysis). *EP300* and *JAG2* were identified as Notch signalling pathway genes that are under constraint for LoF variants. *APLF* is a gene in the DNA repair pathway that harboured a shared LoF variant in 2 affected members of a single family.

The Notch signalling pathway is an intercellular signalling process that is important in embryonic development (Artavanis-Tsakonas et al., 1999) and thought to be perturbed in CRC (Vinson et al., 2016). *EP300* is a histone acetyltransferase that interacts with the intracellular domain of the Notch receptor and its interacting proteins to potentiate transcriptional activation of chromatin (Wallberg et al., 2002). *JAG2* encodes one of 5 canonical ligands that bind to Notch pathway receptors. Somatic variants in *EP300* and *JAG2* occur in approximately 8% and 6% of colon cancers samples in TCGA (TCGA Research Network, 2017) respectively. According to the Cancer Gene Census, somatic *EP300* mutations are found in colorectal, breast, pancreatic and haematological cancers (Forbes et al., 2017). As Notch signalling is thought to be oncogenic in colorectal mucosa due to inhibition of goblet cell differentiation (Katoh and Katoh, 2007), it is unlikely that loss of EP300 predisposes to CRC through its effect on Notch signalling. *EP300* is however also involved in multiple other processes, including cell proliferation (Yao et al., 1998) and stabilisation of *TP53* in response to DNA damage (Yuan et al., 1999). The effect of loss of *JAG2* on Notch signalling in the intestine is not known. *EP300* is on chromosome arm 22q, and *JAG2* is on chromosome arm 14q. These chromosome arms are frequently deleted in CRC (The Cancer Genome Atlas, 2012).

*APLF* is an endo/exonuclease that is important in non-homologous end joining repair of double strand breaks (Macrae et al., 2008). It is also a histone chaperone regulated by poly(ADP-ribose) (Mehrotra et al., 2011). Somatic mutations in APLF are reported in ~3% of rectal adenocarcinoma cases in TCGA, which is the second highest mutation rate of any tumour type in TCGA, behind endometrial cancers at 5%. Interestingly, the family that harboured the *APLF* LoF variant was affected by rectal adenocarcinoma.

No germline syndromes have been attributed to *JAG2* or *APLF* but heterozygous pathogenic variants in *EP300* have been found in up to 8% of patients with Rubinstein-Taybi syndrome, a very rare congenital disorder characterised by growth delay, dysmorphism and intellectual disability, all of which may be mild (Negri et al., 2015). Rubinstein-Taybi syndrome is most commonly caused by a pathogenic variant in *CREBBP* that is found in up to 50% of cases (Bartsch et al., 2005). The incidence of cancer of neural crest origin in children with Rubinstein-Taybi syndrome is approximately 5%, however the molecular subtype of Rubinstein-Taybi syndrome associated with these cancers is unknown. 3 cases, a germ cell tumour (Butler et al., 2016), hepatoblastoma (Milani et al., 2016) and medulloblastoma (Bourdeaut et al., 2014) in Rubinstein-Taybi patients with *CREBBP* mutations have been reported, however cancer cases in Rubinstein-Taybi syndrome patients with *EP300* mutations have not been reported to date. The total number of reported patients with germline *EP300* mutations is low, with only 32 individuals reported in the public Leiden Open Variation Database (Fokkema et al., 2011, Fokkema, 2011). Morphology data is not available for PC16 who had an *EP300* LoF variant in our dataset.

*INPP5E* was found to have a private shared missense variant in 2 affected individuals from a single family (PC2, PC3), as well as an additional LoF variant in an unrelated individual (PC10). *INP55E* is an inositol polyphosphate 5-phosphatase that hydrolyses the lipid inositol phosphates phosphatidylinositol 3,4,5-triphosphate and phosphatidylinositol 4,5-biphosphate (Kisseleva et al., 2000). *INPP5E* loss compromises the spindle assembly checkpoint and results in aneuploidy (Sierra Potchanant et al., 2017b). Biallelic mutations of *INPP5E* account for 3% of Joubert syndrome, a very rare genetically heterogeneous ciliopathy associated with nephronophthisis, mental retardation, ataxia, retinal disease and polydactyly (Hildebrandt et al., 2011, Bachmann-Gagescu et al., 2015). While a proportion of individuals die in childhood, the average survival time is unknown (Dempsey et al., 2017). A case of Burkitt lymphoma in a child with Joubert syndrome of unknown genetic aetiology has been reported (Brinkman et al., 2005). *INPP5E* variants were rare in cancers, with colon cancer having the highest proportion at ~4% in TCGA.

### 4.4.6. *BRCA2* and CRC risk

*A BRCA2* frameshift variant was found in PC10, a male who developed a microsatellite stable rectal adenocarcinoma at the age of 23 years (Table 4.15). PC10 had no family history of cancer, in particular there was no family history of breast, ovarian, pancreatic or prostate cancer. The pedigree of PC10 is shown in Figure 4.6. This is a small family with few women, especially on the father's side, that may explain the lack of a family history, although other rare explanations are a de novo *BRCA2* mutation (Golmard et al., 2016) or non-paternity. This finding was reported back to the referring FCC for diagnostic validation and standard management of cancer risk.

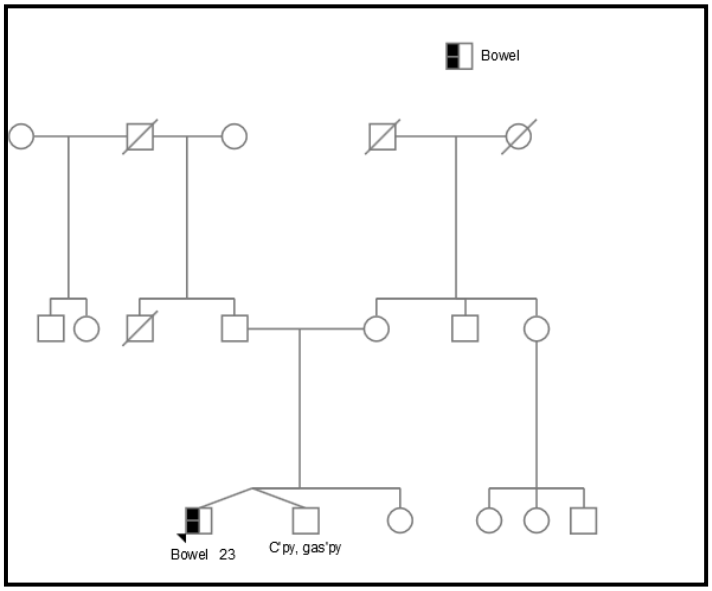



**Figure 4-8. Pedigree of PC10, proband with a *BRCA2* LoF variant, showing lack of a family history of *BRCA2*-related cancers.**

#### *4.4.6.1. Case control study of BRCA2 mutations in early onset familial CRC*

A case control study comparing 847 patients with familial early-onset CRC and 1644 individuals without cancer found no significant increase in *BRCA2* mutations in cases compared with controls (Dobbins et al., 2016). Cases were diagnosed with CRC at or before the age of 55 years, had at least one FDR with CRC, and did not have mutations in known CRC-predisposing genes, while controls were born in 1958 in the United Kingdom. 5 cases and 3 controls were found to have *BRCA2* mutations. As our patient would not have been eligible for inclusion into the cases in this study due to a lack of a family history, a case control study of all patients with early onset CRC, regardless of family history, would be of considerable interest.

#### *4.4.6.2. CRC risk in patients with a BRCA mutation*

Almost all studies evaluating the CRC risk in patients with a known or suspected *BRCA* mutation have been retrospective, with conflicting findings (Sopik et al., 2015). A large prospective cohort study of 7015 women who were *BRCA* mutation carriers (*BRCA1* mutation (n=5481); *BRCA2* mutation (n=1474)*; both BRCA1 and BRCA2* mutations *(n=60))* reported an increased risk of CRC of almost 5 times above the estimated population rate standardised for country and age for *BRCA1* carriers between the ages of 30-49 years (Phelan et al., 2014). Eight CRCs were observed in these younger women with a *BRCA*1 mutation compared to an

expected rate of 1.68. *BRCA2* mutation carriers and those more than 50 years of age showed no evidence for an increased risk. This was a sizable study, although individuals with a *BRCA2* mutation made up just over 20% of study individuals. Limitations of this study include the relatively short mean follow-up of 5 years, the absence of pathology reports for 4 of the 8 CRC cases, the unknown mutation status of known CRC-predisposing genes in the women who developed CRC, and the exclusion of women with ovarian cancer and men from the study. Interestingly, due to the classification of anal cancers as CRC in the control population, 2 anal squamous cell carcinomas in young *BRCA* patients were classified as CRC even though anal cancers are of a different histological subtype and may have arisen from an environmental aetiology. While rare cases of pancreatic squamous cell carcinoma have been reported in patients with an inherited pathogenic variant in *BRCA* (Schultheis et al., 2014, Holter et al., 2015), it is still unclear if squamous cell carcinomass form part of the tumour spectrum for inherited pathogenic variants in *BRCA*. As pointed out in a review of CRC risk in *BRCA* mutation carriers (Sopik et al., 2015), a 5-fold increase in CRC risk to age 50 years gives an estimated risk of 1% by age 50, suggesting that even if *BRCA1* mutations predispose to CRC, they are not very penetrant with regards to CRC risk.

#### *4.4.6.3. LOH and segregation of BRCA2 mutations in familial CRC*

Segregation of *BRCA2* mutations have been reported in 2 familial CRC cases. A case study of 2 sisters who developed CRC before the age of 35 years identified biallelic *BRCA2* mutations in both affected individuals consisting of a frameshift mutation (c.1845_1846delCT, p.Asn615Lysfs*6) and a missense variant (c.7802A4G, p.Tyr2601Cys ) (Degrolard-Courcet et al., 2014). There were no classical clinical features suggestive of Fanconi's anaemia. The older sister developed diffuse large B cell lymphoma at age 31 that was treated with chemotherapy, prior to the development of 3 synchronous CRCs at age 31 and a metachronous rectal cancer at age 35 years. The younger sister developed CRC at age 27 years and breast cancer at age 29, leading to testing and identification of biallelic *BRCA2* mutations. No pathogenic variants were identified in *MLH1, MSH2, MSH6*, *MUTYH*, *TP53* or *BRCA1*. A dizygotic twin of the younger sister died from acute myeloid leukaemia in childhood. An unaffected younger sister was found to have the frameshift *BRCA2* mutation but was wildtype for the missense variant. In a separate study, genetic analysis of *BRCA2* in 44 FCCTX families identified a single family with a germline *BRCA2* frameshift mutation (c.3847_3848delGT p.Val1283Lysfs*2) (Garre et al., 2015). Of the 4 CRCs found in 2 consecutive generations, 2 of 3 tested CRCs were found to have the *BRCA2* mutation, while the tumour that did not was from a family member who developed CRC at the age of 79 years. One tumour was tested for LOH and was found to have lost the wildtype *BRCA2* allele. As our patient is the only affected member in his family, segregation could not be performed but tumour LOH analysis was carried out in Chapter 5. Taken together, there is currently no strong evidence that heterozygous *BRCA2* mutations predispose to CRC. *BRCA2* was not taken forward for the validation phase of this study.

### 4.4.7. Comparison of candidate genes with other germline WES studies in familial/young onset CRC

Seventeen other studies that utilised germline WES in familial and/or young onset CRC or colonic polyposis have been published up until the end of 2015 (Table 4.22), when the analyses described in this chapter were completed. These studies performed WES in 1 to 96 individuals (median of 18), with the number of individuals in our discovery cohort (n=54) being the second highest number of individuals studied at that time. Seven of the published WES studies performed were single family studies.

**Table 4-23. Summary of 17 germline WES studies on familial or young onset CRC**

| First Author | Study Cohort | Analysis Method | Performed segregation within family | LOH studies in tumour | Functional studies | Number of patients who underwent WES |
|---|---|---|---|---|---|---|
| **(Gylfe et al., 2013a)** | Familial | Inter-Family | Y | Y | N | 96 (96 families) |
| **(Esteban-Jurado et al., 2014)** | Familial | Pathway Analysis | Y | N | Y | 43 (29 families) |
| **(Derycke et al., 2013)** | Familial | Intra-Family | Y | N | N | 40 (16 families) |
| **(Nieminen et al., 2014)** | Familial | Intra-Family | Y | Y | N | 4 (1 family) |
| **(Li et al., 2013)** | Familial | Intra-Family | Y | N | N | 4 (1 family) |
| **(Rohlin et al., 2014)** | Familial | Intra-Family | Y | Y | N | 4 (1 family) |
| **(Segui et al., 2015)** | Familial | Intra-Family | Y | Y | Y | 3 (1 family) |
| **(Goldberg et al., 2015)** | Familial | Intra-Family | Y | N/A | N | 2 (1 family) |
| **(Hansen et al., 2015)** | Familial/Polyposis | Intra-Family | Y | N | N | 14 (1 family) |
| **(Weren et al., 2015)** | Polyposis | Inter-Family | Y | N/A | Y | 51 (48 families) |
| **(Gala et al., 2014b)** | Polyposis | Inter-Family, Pathway Analysis | N | N | Y | 20 (20 families) |
| **(Palles et al., 2013)** | Polyposis | Inter-Family, Intra-Family | Y | Y | Y | 15 (15 families) |
| **(Ngeow et al., 2015)** | Polyposis | Pathway Analysis | N | N | Y | 1 |
| **(Tanskanen et al., 2015)** | Young Onset | Inter-Family | N | N | N | 22 (22 families) |
| **(Smith et al., 2013)** | Sporadic CRC (18≤35 years) | Pathway Analysis | N | Y | N | 18 (18 families) |
| **(Arora et al., 2015)** | Young Onset/ Familial/ Polyposis | Pathway Analysis | N | N | Y | 25 (25 families) |
| **(Zhang et al., 2015)** | Young Onset/Familial | Inter-Family, Pathway Analysis | N | N | N | 23 (21 families) |

LOH, Loss-of-heterozygosity. Y, Yes. N, No.

The analysis approaches used in the other studies comprised inter-family, intra-family, pathway analysis and/or genes associated with CRC GWAS regions, previous linkage analyses in CRC or other associations with CRC. There was no overlap between the candidate genes identified in this chapter and the other studies and only limited overlap between those various studies. This lack of overlap suggests high genetic heterogeneity of the remaining unexplained CRC families.

When the candidate genes of the 17 studies were compared with each other, 5 candidate genes were identified in 2 or more studies. *POLE* was prioritised in 5 studies, *ATM* was prioritised in 4 separate studies, and *AKR1C4*, *TSC2* and *CCDC18* were prioritised in 2 separate studies.

Of the 5 studies that prioritised *POLE*, one was the original study that identified a recurrent exonuclease domain mutation in *POLE (POLE* p.Leu424Val (NM_006231; c.1270C>G)), firstly in a family with multiple adenomas and CRC, and subsequently in 12 cases and 0 controls in a large replication cohort (Palles et al., 2013). Of the other 4 studies, two single family studies (Hansen et al., 2015, Rohlin et al., 2014) identified novel *POLE* variants within the exonuclease domain, while the other two studies prioritised *POLE* variants for which the pathogenicity is uncertain: a missense variant (*POLE* p.H1810L) outside the exonuclease domain (Esteban-Jurado et al., 2014), and a deletion (ENST00000320574; c.5621_5622delGT) (Smith et al., 2013). LoF variants in *POLE* are not thought to be pathogenic (Briggs and Tomlinson, 2013). Germline missense variants within the exonuclease region of *POLE* have been seen in other CRC studies (Bellido et al., 2016, Valle et al., 2014) and are associated with a hypermutated tumour phenotype (Palles et al., 2013).

*ATM* is a serine-threonine kinase that is involved in the detection of double-strand DNA breaks and cell cycle checkpoint activation. Ataxia telangiectasia is a rare recessive genetic syndrome caused by mutations in *ATM,* most commonly compound heterozygous truncating mutations. The syndrome manifests with nervous system and immune disorders in childhood, including cerebellar ataxia, oculomotor apraxia and immunodeficiency as well as an increased risk of blood cancers. Carriers of an *ATM* mutation are thought to have an increased risk of other cancers, with a systematic review of studies of relatives of ataxia-telangiectasia probands finding an increased risk of breast cancer (van Os et al., 2016).

The *ATM* variants in the 4 studies comprised a homozygous missense variant (Tanskanen et al., 2015) and 3 heterozygous variants (2 truncating and 1 missense) (Smith et al., 2013, Zhang et al., 2015, Esteban-Jurado et al., 2014). CRC risk for heterozygous carriers of *ATM* mutations has not been as well-studied, and studies of relatives of ataxia-telangiectasia probands have not shown a marked increase in CRC risk (Thompson et al., 2005, Olsen et al., 2005, Geoffroy-Perez et al., 2001). *ATM* is known to have a large number of missense variants, which are a confounding factor in association studies of cancer risk (Scott et al., 2002). Hence it is still unclear if variants in *ATM* predispose to CRC, and if so which variants are pathogenic.

A recurrent LoF variant in *AKR1C4* was found in 2 of 96 cases by inter-family analysis (Gylfe et al., 2013b), while a different LoF variant was found in 1 individual in a second study, and

prioritised due its presence within a linkage region associated with CRC (Esteban-Jurado et al., 2014). There was no LOH of the wildtype allele in either study, suggesting that *AKR1C4* is not a classical tumour suppressor gene. Segregation analysis performed in one study (Gylfe et al., 2013b) showed that the variant segregated within the family. *AKR1C4* encodes an enzyme involved in bile acid and steroid metabolism in the liver (Penning et al., 2004).

Two studies (Esteban-Jurado et al., 2014, Zhang et al., 2015) identified different potentially deleterious missense variants in *TSC2*, a gene prioritised through pathway analysis. Segregation and LOH studies were not performed. Mutations in *TSC2* cause tuberous sclerosis, an autosomal dominant multi-organ disease characterised by epilepsy, characteristic skin and/or retinal lesions and benign tumours in the brain and other organs. Rare associations with renal cell carcinoma and brain astrocytoma (Al-Saleem et al., 1998), and pancreatic neuroendocrine tumours (Larson et al., 2012) have been noted in tuberous sclerosis, however an increased incidence of CRC has not been reported.

Two different LoF variants were found in *CCDC18* in one study (Gylfe et al., 2013b) while a potentially deleterious missense variant was identified in another study (Esteban-Jurado et al., 2014). There was no LOH of the wildtype allele in the tumours from either study and segregation was not performed. The function of *CCDC18* is not known.

Apart from *POLE*, further studies are required to determine the role of *ATM, AKR1C4, TSC2* and *CCDC18* in CRC predisposition.

### 4.4.8. Conclusion

The identification of inherited pathogenic variants in PC15, PC50 and PC51 through WES has enabled these individuals and their families to have tailored genetic counselling and appropriate screening aimed at reducing the risk of associated cancers. As WES has become available for use in the lab and in the clinic, the prospect of identifying novel genes predisposing to familial CRC using WES is attractive for patients in whom a pathogenic variant in a known CRC-predisposing gene has not been identified. Through analysis of PPV from WES data in a discovery cohort, 45 novel candidate CRC-predisposing genes were prioritised for validation in a larger cohort of Australian patients with young onset CRC. The genes identified solely in the family with AMTs were not included, as the validation cohort did not include any individuals with AMTs. While there were additional genes that were of interest and additional analyses that could be performed, due to cost constraints the top 45 genes were selected for screening in a further 333 individuals with young onset CRC, using a targeted gene panel (Chapter 5).

# Chapter 5. Validation Studies: Targeted Gene Re-sequencing of Candidate CRC-Predisposing Genes in a Validation Cohort; LOH analysis in Tumour Tissue and Segregation Analysis in the Discovery Cohort

## 5.1. Introduction

Having identified 45 candidate CRC-predisposition genes from the exome sequencing of the discovery cohort, further evidence was sought in a validation cohort that these genes predispose to CRC when mutated. Firstly, genes that predispose to CRC are expected to be mutated in individuals with CRC at a higher frequency than unaffected controls. Secondly, PPVs in CRC-predisposing genes are also expected to segregate with affected individuals within a family. Thirdly, CRC-predisposing genes that function as tumour suppressor genes have loss of the wild type allele in tumour tissue.

This chapter describes 1) the recruitment of a validation cohort of young-onset or familial CRC and the results of targeted re-sequencing of candidate CRC-predisposing genes in this cohort, 2) assessment for LOH of PPVs in CRCs from the individuals in the discovery cohort, and 3) segregation of the PPVs identified in Chapter 4 within families from the discovery cohort, where available.

### 5.1.1. Obtaining a population-matched validation cohort

Consortia such as the VCB (VIC, Australia) and kConFab (Osborne et al., 2000) collect tissue samples, including germline DNA, for cancer research. The VCB is a Victorian government (Victoria, Australia) supported tissue bank that obtains tissue and clinical data from consenting healthy or cancer-affected donors from 27 hospitals in Melbourne, Australia. kConFab collects tissue and clinical data from consenting Australian and New Zealand families with a strong history of breast cancer, with detailed annotation of all cancer types in family members. Such consortia allow access to a population-matched validation cohort.

### 5.1.2. The Haloplex Target Enrichment System for targeted gene sequencing

Targeted gene sequencing platforms are generally either hybridisation or amplification-based (Mertes et al., 2011), while the Haloplex Target Enrichment System (Agilent Technologies, Santa Clara, CA, USA) uses DNA fragment circularisation to select for targeted regions of the genome. This circularisation technique has superior specificity compared to other methods (Mertes et al., 2011). Genomic DNA is first digested by restriction enzymes that provide specific complementary binding sites to the target probes. Digested DNA containing the targeted regions bind to both ends of the DNA probes. Following probe hybridisation the biotinylated probes are pulled down and bound genomic DNA is ligated. The circularised DNA targets undergo PCR amplification prior to sequencing. This technique requires only 200 ng of input DNA.

### 5.1.3. Emerging CRC-predisposing genes

In addition to the candidate CRC-predisposing genes identified in this study and described in Chapter 4, since the start of this work several other studies have proposed other candidate CRC-predisposing genes, based on recurrent mutations in individuals affected with CRC.

#### *5.1.3.1. NTHL1*

WES of 51 individuals with multiple colonic adenomas revealed a recurrent homozygous nonsense mutation in *NTHL1*, a base excision repair gene (Weren et al., 2015). The *NTHL1* NM_002528, c.268C>T, p.Q90* mutation was identified in 7 individuals from 3 unrelated CRC families, and segregated with disease. An excess of somatic C:G>T:A transitions was detected in the corresponding CRCs, consistent with a base excision repair defect . The homozygous mutations in *NTHL1* were also associated with endometrial and duodenal malignancy. A case report of a 41 year old female individual with compound heterozygous LoF *NTHL1* mutations described additional cancers including breast, bladder, head and neck cancers and basal cell carcinoma (Rivera et al., 2015).

#### *5.1.3.2. BUB1B*

An individual with mosaic variegated aneuploidy detected in lymphocytes and fibroblasts developed multiple gastrointestinal cancers, namely an ampulla of Vater cancer at age 34 years, followed by 2 gastric cancers and CRC in their 60s as well as colorectal and gastric adenomas (Rio Frio et al., 2010). A germline homozygous intronic mutation in *BUB1B* was detected, that resulted in the creation of a new preferred splice site. *BUB1B* encodes BUBR1, a component of the mitotic checkpoint complex that prevents completion of mitosis in cells with unaligned chromosomes (Sudakin et al., 2001). BUBR1 also interacts with APC, which may explain the increased risk of GI cancers in this individual (Zhang et al., 2007, Rio Frio et al., 2010).

#### *5.1.3.3. BUB3 and BUB1*

*BUB3* is also a member of the mitotic checkpoint complex. *BUB1* is required for the localisation of the checkpoint proteins BUBR1, CENPE, CENPF and MAD2 to the kinetochore (Johnson et al., 2004). BUB1 is also needed for the correct attachment of kinetochores to microtubules (Meraldi and Sorger, 2005).

A 1.75 Mb deletion of chromosome 2q13 that includes *BUB1* was detected in an individual who had developed CRC at the age of 37 years (de Voer et al., 2013). Variegated aneuploidies and structural abnormalities were detected in the individual's lymphocytes and fibroblasts. WES of a further 33 patients of Dutch or Chinese descent who developed CRC prior to 40 years of age, identified 2 heterozygous germline LoF mutations in *BUB1* and a heterozygous missense mutation that was highly conserved in *BUB3* . Targeted sequencing and copy number analysis of 174 familial or young onset CRC cases identified 2 further missense variants in *BUB3* that were not detected in 1154 control cases. Two of the 3 missense variants (NM_004725.3:c.63G>C and NM_004725.3:c790T>C) were predicted to be pathogenic or possibly damaging by in silico methods, and variegated aneuploidies and structural abnormalities were detected in the fibroblasts and lymphocytes of the affected individuals. One of the individuals with a LoF *BUB1* mutation also carried a pathogenic germline *MLH1*

variant (c.453+1G>T), although interestingly the individual's sister who carried the *MLH1* mutation but not the *BUB1* mutation had not developed CRC by age 58 years.

#### *5.1.3.4. KMT2C*

In a single family affected by young onset CRC and AML, an insertion in *KMT2C (MLL3)* was detected in 4 affected individuals, 2 of whom developed CRC at the ages of 43 and 59 years, while the other 2 developed AML in their 40s (Li et al., 2013). KMT2C is a histone methyltransferase and has a high LoF variant constraint score (pLI score) in the ExAC dataset, of 1.0.

### 5.1.4. Aims and Objectives

The aim of this chapter is to seek further evidence for CRC predisposition of the candidate CRC-predisposing genes identified in Chapter 4 (n=45) or proposed in the literature (n=5, Section 5.1.3) by re-sequencing these genes in a validation cohort. For the candidate genes identified in Chapter 4, two further aims are assessing the corresponding CRCs for somatic LOH of the PPV in the candidate gene and determining if the PPV segregates with disease within the family.

The specific objectives of the work described in this chapter are to:

a) recruit a validation cohort of patients with available germline DNA for targeted gene re-sequencing of candidate CRC-predisposing genes

b) identify PPVs in candidate CRC-predisposing genes in the validation cohort using the Haloplex Targeted Enrichment System (Agilent Technologies, Santa Clara, CA, USA)

c) compare the frequency of LoF variants in candidate CRC-predisposing genes to that of the non-TCGA subset of the ExAC dataset

d) determine if the wild type allele of the PPV has been lost in CRCs from the discovery cohort patients in whom a candidate CRC-predisposing gene was identified

e) perform segregation analysis of PPV in available family members of index cases from the discovery cohort

## 5.2. Methods

### 5.2.1. Design and Filtering of Haloplex Data

#### *5.2.1.1. Design*

The genes on the custom Haloplex Targeted Enrichment System included the candidate CRC-predisposing genes identified in Chapter 4 (n=45), known accepted CRC genes, and emerging CRC-predisposing genes from Section 5.1.3 (*NTHL1, BUB1B, BUB1, BUB3* and *KMT2C*) (n=5). Known CRC genes (*APC, MLH1, MSH2, MSH6, PMS2, MUTYH, STK11, SMAD4, BMPR1A, POLE, POLD1*) were included on the targeted sequencing panel as the VCB and kConFab individuals may not have been screened for mutations in known CRC-predisposing genes, in contrast to individuals recruited from FCCs. *BRCA2* was omitted from the targeted sequencing panel as it has already been studied extensively (discussed in Chapter 4 Section 4.4.6). The final list of genes on the custom Haloplex Targeted Enrichment System kits are summarised in Table 5.1.

**Table 5-1. List of genes on custom Haloplex Targeted Enrichment System kit used for sequencing of validation cohort**

| Gene | Analysis method |
|---|---|
| *ASCL3* | inter-family LoF analysis |
| *GALC* | inter-family LoF analysis |
| *GALNT8* | inter-family LoF analysis |
| *GPR144* | inter-family LoF analysis |
| *MS4A15* | inter-family LoF analysis |
| *PDE5A* | inter-family LoF analysis |
| *RIPPLY3* | inter-family LoF analysis |
| *SERPINB7* | inter-family LoF analysis |
| *SH2D4A* | inter-family LoF analysis |
| *TEX14* | inter-family LoF analysis |
| *TNFRSF10B* | inter-family LoF analysis |
| *ZNF788* | inter-family LoF analysis |
| *NEDD4* | inter-family LoF analysis |
| *TIMELESS* | inter-family LoF analysis |
| *FKBP10* | intra-family missense analysis |
| *GOLGA5* | intra-family missense analysis |
| *GPR137B* | intra-family missense analysis |
| *HCN3* | intra-family missense analysis |
| *HTR5A* | intra-family missense analysis |
| *IL4R* | intra-family missense analysis |
| *INPP5E* | intra-family missense analysis |
| *LRRK2* | intra-family missense analysis |
| *MMP2* | intra-family missense analysis |
| *PLSCR3* | intra-family missense analysis |
| *SF3A3* | intra-family missense analysis |
| *SUCLG2* | intra-family missense analysis |
| *TMCC3* | intra-family missense analysis |
| *UFL1* | intra-family missense analysis |
| *CNTN6* | CRC pathway LoF analysis |
| *DMC1* | CRC pathway LoF analysis |
| *ERCC3* | CRC pathway LoF analysis |
| *GDF2* | CRC pathway LoF analysis |
| *LRRFIP2* | CRC pathway LoF analysis |
| *MUS81* | CRC pathway LoF analysis |
| *NOTCH4* | CRC pathway LoF analysis |
| *POLN* | CRC pathway LoF analysis |
| *USP28* | CRC pathway LoF analysis |
| *WNT10A* | CRC pathway LoF analysis |
| *XRCC6BP1* | CRC pathway LoF analysis |

LoF, loss-of-function.

**Table 5-1 (continued). List of genes on custom Haloplex Targeted Enrichment System kit used for sequencing of validation cohort**

| Gene | Analysis Method |
|---|---|
| ***INCENP*** | Gene constraint LoF analysis |
| ***PHRF1*** | Gene constraint LoF analysis |
| ***SIN3B*** | Gene constraint LoF analysis |
| ***EP300*** | CRC pathway LoF analysis; gene constraint LoF analysis |
| ***JAG2*** | CRC pathway LoF analysis; gene constraint LoF analysis |
| ***APLF*** | CRC pathway LoF analysis; intra-family LoF analysis |
| ***BUB1*** | CRC-predisposing gene from literature |
| ***BUB1B*** | CRC-predisposing gene from literature |
| ***BUB3*** | CRC-predisposing gene from literature |
| ***KMT2C*** | CRC-predisposing gene from literature |
| ***NTHL1*** | CRC-predisposing gene from literature |
| ***APC*** | Known CRC-predisposing gene |
| ***BMPR1A*** | Known CRC-predisposing gene |
| ***MLH1*** | Known CRC-predisposing gene |
| ***MSH2*** | Known CRC-predisposing gene |
| ***MSH6*** | Known CRC-predisposing gene |
| ***MUTYH*** | Known CRC-predisposing gene |
| ***PMS2*** | Known CRC-predisposing gene |
| ***POLD1*** | Known CRC-predisposing gene |
| ***POLE*** | Known CRC-predisposing gene |
| ***SMAD4*** | Known CRC-predisposing gene |
| ***STK11*** | Known CRC-predisposing gene |

LoF, loss-of-function.

#### *5.2.1.2. Filtering*

Variants identified by at least one of the three variant callers (GATK Unified Genotyper, GATK Haplotype Caller or Platypus) were considered for further analysis. Analysis was limited to LoF variants, missense variants and in-frame indels.

##### 5.2.1.2.1. Population frequencies

For known CRC-predisposing genes with an autosomal dominant mode of inheritance (*APC, MLH1, MSH2, MSH6, PMS2, STK11, BMPR1A, SMAD4, POLE, POLD1*), LoF and missense variants with a population allele frequency of <0.5% in the 1000 Genomes Project, ESP or Non-TCGA Subset of EXAC were examined for pathogenicity. For known CRC-predisposing genes with a recessive mode of inheritance (*MUTYH, NTHL1, BUB1B*), compound heterozygous or homozygous LoF and missense variants with a population allele frequency of<5% in the 1000 Genomes Project, ESP or Non-TCGA Subset of TCGA were examined for pathogenicity. Variants in the candidate CRC-predisposing genes had to have a population allele frequency of <0.1% in 1000G, ESP and the Non-TCGA Subset of ExAC.

#### 5.2.1.2.2. Quality filters

Variants with a rare population frequency but observed in more than 5% of the 337 CRC samples in the validation cohort plus 3 control samples (sequenced as part of the technical requirement of the Haloplex platform) were excluded, as these are likely to represent sequencing or alignment artefacts.

Further quality filters were applied using GATK FisherStrand and ReadPosRankSum annotations. The FisherStrand value uses Fisher's Exact Test to determine if the alternate allele was seen more or less frequently on the forward or reverse allele than the reference strand. This value is Phred-scaled, with a higher value indicating strand bias and reflecting a false positive call. The ReadPosRankSum test gives an indication of whether or not the alternate allele is seen more or less frequently at the ends of sequencing reads compared to the reference allele, as sequencing errors are most common at the ends of reads. This is derived using the z-score of the Mann-Whitney-Wilcoxon rank sum test for the positions of the alternate and reference alleles on reads. When the alternate allele is seen at the ends of reads more often than the reference allele, the ReadPosRankSum test returns a negative score. Variants with a FisherStrand value >60 or a ReadPosRankSum test score of <-8 were omitted as per GATK recommendations released in 2016 (Broad Institute, 2016).

Variants with an alternate allele frequency of <0.3 were omitted as heterozygous variants were expected to have an alternate allele frequency close to 0.5, and homozygous variants were expected to have an alternate allele frequency close to 1.0. Variants that were unidirectional and with a QUAL score <100 were also omitted.

### 5.2.2. Evaluating the frequency of LoF variants in a gene in the Non-TCGA subset of the ExAC database database

The Non-TCGA subset of the ExAC database was used to estimate the expected frequency of LoF variants for a given gene in a control cohort. For every candidate CRC-predisposing gene in which a LoF variant was detected in the validation cohort, the frequency of LoF variants in the corresponding transcript was examined in the Non-TCGA subset of the ExAC database. The frequency of LoF variants in a gene transcript in the Non-TCGA subset of the ExAC database was calculated using the sum of all LoF alleles for the corresponding transcript. This sum was divided by the expected number of alleles sequenced in the Non-TCGA subset of the ExAC database (n=106210), multiplied by 200. This gives an estimate of the percentage of individuals in the Non-TCGA subset of the ExAC database who carry a LoF variant in a gene, as each individual carries 2 alleles.

The frequency of rare LoF variants in the validation dataset was calculated using the sum of all rare LoF variants in a gene, according to transcript, divided by the number of alleles for the gene in the validation dataset from individuals without a known CRC-predisposing gene mutation, multiplied by 200.

### 5.2.3. Assessing CRCs for LOH of PPV

Letters were sent to pathology service providers requesting tumour blocks from resected CRCs of probands from the discovery cohort, in whom a PPV in a candidate CRC-predisposing gene had been identified. Ten 5 µm sections were cut from each tumour block by the Pathology Department at PMCC and mounted onto glass slides. Tumour DNA was extracted as described in Chapter 2 Section 2.2.10.

### 5.2.4. Segregation of PPV identified in Chapter 4 within Families from the Discovery Cohort

Letters inviting relatives of the probands to participate in this study were sent to each proband to pass on to affected and unaffected family members. Relatives who provided contact details in response to the invitation letter were contacted and the study was explained. Relatives who consented to participate in the study were sent a Patient Information and Consent Form and a copy signed by the relative was returned to the study coordinator. A pathology slip was also sent for blood collection in order to obtain germline DNA. Sequencing of the PPVs identified in Chapter 4 was performed on germline DNA obtained from relatives who consented to the study.

## 5.3. Results

### 5.3.1. Characteristics of the validation cohort

Germline DNA samples for a validation cohort were sought from tissue biobanks as well as by recruitment through the FCCs as previously described in Chapter 3. The sample breakdown of the final validation cohort is summarised in Table 5.2.

**Table 5-2. Sources of germline DNA for validation cohort**

| Source | Number of samples | Criteria |
|---|---|---|
| VCB | 133 | CRC≤50 years of age |
| FCCs | 117 | See section 5.2.1 |
| MCO | 55 | CRC between 41-50 years of age |
| kConFab | 32 | CRC≤50 years of age |
| **Total** | **337** | |

VCB, Victorian Cancer Biobank. FCCs, Familial Cancer Centres. MCO, Molecular and Cellular Oncology Group. kConFab, Kathleen Cuningham Foundation.

### 5.3.2. Performance summaries of Targeted Gene Sequencing Panel

The performance summaries for targeted gene sequencing of 337 samples using the Haloplex Targeted Enrichment System is shown in Appendix 2, Table 1. Between 1.58 million and 7.88 million reads were generated per sample. On average 99% of reads were mapped to the reference sequence. The percentage reads on target ranged from 86% to 93%, with a mean of 89%. The percentage of target bases covered at least 10 times was at least 91%, with a mean of 96%. The percentage of target bases covered at least 20 times was at least 86%, with a mean of 95%. The mean coverage for target bases per sample ranged from 188 times to 961 times, with a mean of 358 times.

### 5.3.3. Analysis of LoF and missense variants in known CRC-predisposing genes

#### *5.3.3.1. APC*

Three individuals carried the *APC* I1307K variant (Table 5.3) that is currently known to be pathogenic (Laken et al., 1997, Gryfe et al., 1999).

#### *5.3.3.2. MLH1*

A 4 bp deletion at an essential splice site was detected in *MLH1* in 1 individual (Table 5.3). The genomic sequence at the intron exon boundary where the deletion occurs consists of a duplication of AGAA, hence deletion of AGAA does not change the essential splice site, but results in a frameshift, with a predicted stop codon 60 base pairs downstream.

**Table 5-3. Pathogenic variants in *APC* and *MLH1* in 4 individuals from the validation cohort**

| Gene | Sample | Mutation type | Ensembl Transcript | cDNA variant | Amino acid change | Variant Frequency |
|---|---|---|---|---|---|---|
| *APC* | CRC-NSW-134275<br>CRC-VCB-10-00793<br>CRC-FCC-5507 | missense | ENST00000257430.4 | c.3920T>A | p.Ile1307Lys | 0.42-0.45 |
| *MLH1* | CRC-VCB-07RMH049 | deletion | ENST00000231790.2 | c.208-2_209delAGAA | N/A | 0.58 |

#### *5.3.3.3. POLE*

Inspection of the exonuclease region of *POLE* identified 1 rare missense variant ENST00000320574.5:c.1085A>G, p.Y362C, in 1 individual (CRC-4273-000). This variant is of uncertain significance in ClinVar (National Center for Biotechnology Information, Landrum et al., 2018). The variant is predicted to be deleterious by Condel, and has a CADD scaled score of 26.1, placing it in the top 1% most deleterious substitutions in the human genome. This individual did not have colonic polyps.

#### *5.3.3.4. POLD1*

Inspection of the exonuclease region of *POLD1* identified 1 rare missense variant ENST00000440232.2:c.1439A>T, p.H480L, in 1 individual (CRC-VCB-10PM02139). This variant has not been reported in the literature. The variant is predicted to be neutral in Condel, and has a CADD scaled score of 18.95. A right hemicolectomy specimen from this individual did not contain any polyps.

#### *5.3.3.5. MUTYH*

All *MUTYH* missense mutations identified were heterozygous, and there were no instances of compound heterozygosity in the validation cohort. The two most common pathogenic missense mutations in the European population, NM 001128425.1:c.1187G>A, p. G396D and NM 001128425.1:c.536A>G, p.Y179C, were detected in 5 individuals and 1 individual, respectively. Another individual carried the NM 001128425.1.c.1118C>T, p.A373V mutation, that had been described in the literature in a Korean patient with compound heterozygous biallelic MUTYH pathogenic variants, rectal cancer and 36 polyps at the age of 39 years (Kim et al., 2007).

#### *5.3.3.6. Other known CRC-predisposing genes*

There were no rare LoF variants in *MLH2, MLH6*, *PMS2, STK11, BMPR1A or SMAD4.* The 4 samples with *APC* or *MLH1* mutations were omitted from further analysis.

### 5.3.4. Analysis of candidate CRC-predisposing genes

#### *5.3.4.1. LoF variants in candidate CRC-predisposing genes in the validation cohort*

Fifty candidate CRC-predisposing genes were analysed in the remaining 333 individuals from the validation cohort. Only 14 of the 50 candidate CRC-predisposing genes had 1 or more LoF variants identified in the validation cohort (Table 5.4) and all were heterozygous. No compound heterozygous or homozygous LoF variants were identified in *NTHL1* or *BUB1B*, genes with an autosomal recessive mode of inheritance. Seven genes had a LoF variant in 2 or more samples. All LoF variants were identified in the canonical transcript except as described.

*ZNF788* had a LoF variant identified in 5 samples. All LoF variants were in the last exon (exon 3). Subsequent to the prioritisation of *ZNF788* in the discovery cohort and re-sequencing in the validation cohort, *ZNF788* has been annotated as a possible pseudogene product by the Universal Protein Resource (The UniProt Consortium, 2018) and hence was not considered further.*WNT10A* had a LoF variant in 3 samples. Five genes (*NEDD4, GPR144, MS4A15, TNFRSF10B*, *LRRK2*) had a LoF variant in 2 samples. One of the 2 LoF variants identified for *NEDD4* was only seen in the non-canonical transcript, and was located outside the genomic coordinates of the canonical transcript, and hence was not annotated on the canonical transcript. One of the 2 LoF variants in *GPR144* was in the last exon (exon 20), hence it may not be deleterious as it is not predicted to cause nonsense-mediated decay.

Seven genes (*GALNT8, ASCL3, ERCC3, NOTCH4, FKBP10, INPP5E, UFL1*) had 1 LoF variant identified in the validation cohort. For *ASCL3, FKBP10* and *INPP5E*, the LoF variant was in the last exon (exon 2, 10 and 10 respectively).

Sequencing data for the large number of individuals sequenced in the non-TCGA subset of the ExAC dataset (n=53 105) was leveraged to give an approximation of the percentage of individuals with a LoF variant in a control population. The estimated percentage of individuals with a LoF variant in the non-TCGA subset of the ExAC dataset is shown in Table 5.4. The mean sequencing coverage of candidate CRC-predisposing genes in the ExAC database was taken into account (table 5.4). *GPR144* had a mean coverage of less than 20X, which may result in an underestimation of the percentage of individuals with a LoF variant in the non-TCGA subset of the ExAC dataset.

The six genes with multiple LoF variants in the validation cohort appear to have an increased incidence of LoF variants compared to the estimated frequency of LoF variants in the non-TCGA subset of the ExAC dataset (Table 5.4). For *GPR144,* as the sequence coverage is less than 20X in ExAC, the difference in observed rates of LoF between the validation cohort and the non-TCGA subset of ExAC could be due to inadequate coverage in ExAC., Of the seven genes with one LoF variant seen in the validation cohort, the percentage of LoF variants in the validation cohort approximates that seen in the non-TCGA subset of the ExAC dataset for 5 genes (*GALNT8, ASCL3, ERCC3, NOTCH4* and *FKBP10*) hence they are unlikely to be CRC-predisposing genes. The LoF variant in *INPP5E* is in the last exon (exon 10), hence it may not be deleterious as it is not predicted to cause nonsense-mediated decay.

#### *5.3.4.2. Novel missense variants identified from intra-family analysis of the discovery cohort*

Fourteen missense variants not seen in the non-TCGA subset of ExAC (*i.e.* novel) were identified through intra-family analysis of the discovery cohort. There were no recurrences of the shared private missense variants identified from the intra-family analysis in unrelated individuals in the validation cohort. For *GOLGA5, HTR5A, and FKBP10,* a different novel missense variant with a CADD score of>20 was identified (Table 5.5). While these high CADD scores imply pathogenicity, the numbers are too small to determine if these variants are truly CRC-predisposing or if these novel missense variants have occurred by chance.

**Table 5-4. LoF variants in the validation cohort**

| Gene | No. of samples with LoF variant | Sample | Ensembl Transcript | cDNA | Amino acid change | Mean coverage in ExAC | % of individuals in validation dataset with a LoF variant | Estimated % of individuals in non-TCGA subset of ExAC with LoF variant |
|---|---|---|---|---|---|---|---|---|
| ***WNT10A*** | 3 | CRC-NSW-11555 | ENST00000258411.3 | c.321C>A | p.Cys107Ter | 36.98 | 0.9 | 0.13 |
| | | CRC-4083-000 CRC-KCF-01-005-0555 | ENST00000258411.3 | c.376+1G>T | . | | | |
| ***MS4A15*** | 2 | CRC-VCB-10CH094 | ENST00000405633.3 | c.612+1G>C | . | 52.00 | 0.59 | 0.03 |
| | | CRC-MMC-5307 | ENST00000405633.3 | c.372_384delAGTGGCAGCCGAG | p.Val125ArgfsTer8 | | | |
| ***TNFRSF10B*** | 2 | CRC-VCB-10SH156 | ENST00000276431.4 | c.598delT | p.Ser200ProfsTer17 | 27.95 | 0.59 | 0.14 |
| | | CRC-VCB-05PM0609 | ENST00000276431.4 | c.748+1G>A | . | | | |
| ***LRRK2*** | 2 | CRC-KCF-00-009-0341 CRC-KCF-04-005-0715 | ENST00000298910.7 | c.1645_1646insAGAT | p.Ala549GlufsTer28 | 57.46 | 0.59 | 0.23 |
| ***NEDD4*** | 2 | CRC-VCB-09PM0241 | ENST00000338963.2 | c.1147C>T | p.Arg383Ter | 67.02 | 0.29 | 0.08 |
| | | CRC-VCB-06RMH224 | ENST00000435532.3* | c.291+1G>A | . | 51.44 | 0.29 | 0.04 |
| ***GPR144*** | 2 | CRC-3950-000 | ENST00000334810.1 | c.2860_2861insC | p.Arg956ProfsTer? | 6.88 | 0.59 | 0.19 |
| | | CRC-WA-5561 | ENST00000334810.1 | c.1681C>T | p.Gln561Ter | | | |
| ***UFL1*** | 1 | CRC-2958-000 | ENST00000369278.4 | c.171_172insA | p.Thr58AsnfsTer6 | 55.22 | 0.29 | 0.11 |
| ***INPP5E*** | 1 | CRC-NSW-13982 | ENST00000371712.3 | c.1885C>T | p.Gln629Ter | 28.08 | 0.29 | 0.01 |
| ***NOTCH4*** | 1 | CRC-3938-002 | ENST00000375023.3 | c.4927C>T | p.Arg1643Ter | 52.63 | 0.29 | 0.37 |

LoF, loss-of-function, ExAC, Exome Aggregation Consortium.

*non-canonical transcript

**Table 5-4 (continued). LoF variants in validation cohort**

| *Gene* | No. of samples with LoF variant | Sample | Ensembl Transcript | cDNA | Amino acid change | Mean coverage in ExAC | % of individuals in validation dataset with a LoF variant | Estimated % of individuals in non-TCGA subset of ExAC with LoF variant |
|---|---|---|---|---|---|---|---|---|
| ***FKBP10*** | 1 | CRC-VCB-10PM1612 | ENST00000321562.4 | c.1621C>T | p.Gln541Ter | 55.83 | 0.29 | 0.30 |
| ***ASCL3*** | 1 | CRC-VCB-2009-00711 | ENST00000325884.1 | c.475_476dupAA | p.Asn159LysfsTer9 | 85.06 | 0.29 | 0.27 |
| ***GALNT8*** | 1 | CRC-4271-000 | ENST00000252318.2:c.1357C>T | c.1357C>T | p.Gln453Ter | 58.86 | 0.29 | 0.24 |
| ***ERCC3*** | 1 | CRC-4356 | ENST00000285398.2 | c.325C>T | p.Arg109Ter | 67.79 | 0.29 | 0.22 |

LoF, loss-of-function, ExAC, Exome Aggregation Consortium.

**Table 5-5. Novel missense variants in the validation cohort, in genes with shared novel missense variants from the discovery cohort**

| **Validation Cohort** | | | | | | **Discovery cohort** | | |
|---|---|---|---|---|---|---|---|---|
| Gene | Sample | Ensembl transcript | cDNA | amino acid change | CADD Phred Score | Ensembl transcript | amino acid change | CADD Phred Score |
| ***GOLGA5*** | CRC-MMC-5272 | ENST00000163416.2 | c.1207G>A | p.Glu403Lys | 25.8 | ENST00000163416.2 | p.Asp681Asn | 34 |
| ***HTR5A*** | CRC-NSW-16855 | ENST00000287907.2 | c.454C>T | p.Arg152Cys | 21.7 | ENST00000287907.2 | p.Asp121Asn | 27.5 |
| ***FKBP10*** | CRC-VCB-11PM1483 | ENST00000321562.4 | c.326T>C | p.Met109Thr | 21.3 | ENST00000321562.4 | p.Ser419Leu | 27.5 |

### 5.3.5. Investigation of LOH of the PPV in tumour tissue

Of the 31 individuals with a PPV in a candidate CRC-predisposing gene (identified in Chapter 4), 30 had CRC. One individual (PC54) had 47 colonic polyps removed over time without developing CRC. Tumour blocks or 10 unstained slides from 30 CRCs were requested from 13 pathology departments.

Twenty CRCs were received for micro-dissection. For PC6, a metachronous CRC at age 70 was received, however the CRC from age 40 could not be located. CRCs from 10 individuals had either been destroyed or were not available.

There was <25% tumour purity in very small tumours in 5 tumour blocks, these were excluded. Tumour was micro-dissected from slides of the other tumour blocks, and DNA was extracted. Insufficient DNA was obtained from 1 sample. Primers were designed for PCR products of between 83 and 130 base pairs, targeting the region containing the PPV. Despite multiple attempts, Sanger sequencing was only successful for 4 variants (Table 5.6).

**Table 5-6. Summary of Sanger sequencing findings for PPV in tumour tissue**

| Sample | Gene | PPV | LOH |
|---|---|---|---|
| **PC6** | *FKBP10* | NM_021939.3: c.1256C>T | No |
| **PC7** | *FKBP10* | NM_021939.3: c.1256C>T | No |
| **PC17** | *ASCL3* | NM_020646.1: c.382C>T | Yes – homozygous wildtype |
| **PC36** | *JAG2* | NM_145159.2: c.3314_3315insG | No |

The loss of the mutant allele of *ASCL3* in PC17 indicates that the PPV in *ASCL3* is not a CRC-predisposing gene mutation. No LOH was identified in the other 3 samples, which does not preclude *FKBP10* and *JAG2* from being candidate CRC-predisposing genes, as a different somatic mutation or gene methylation may inactivate the second allele, or the PPV may have a dominant negative effect.

### 5.3.6. Segregation of PPVs within families with multiple affected family members

Of the 26 families from which a candidate CRC-predisposing gene had been prioritised, 13 families had additional affected family members. Six families were from the MCO cohort from which DNA from additional family members was not available. Of the remaining 7 families, an affected individual from 1 family agreed to participate in the study, while individuals from the remaining 6 families did not respond to the invitation to participate in the study.

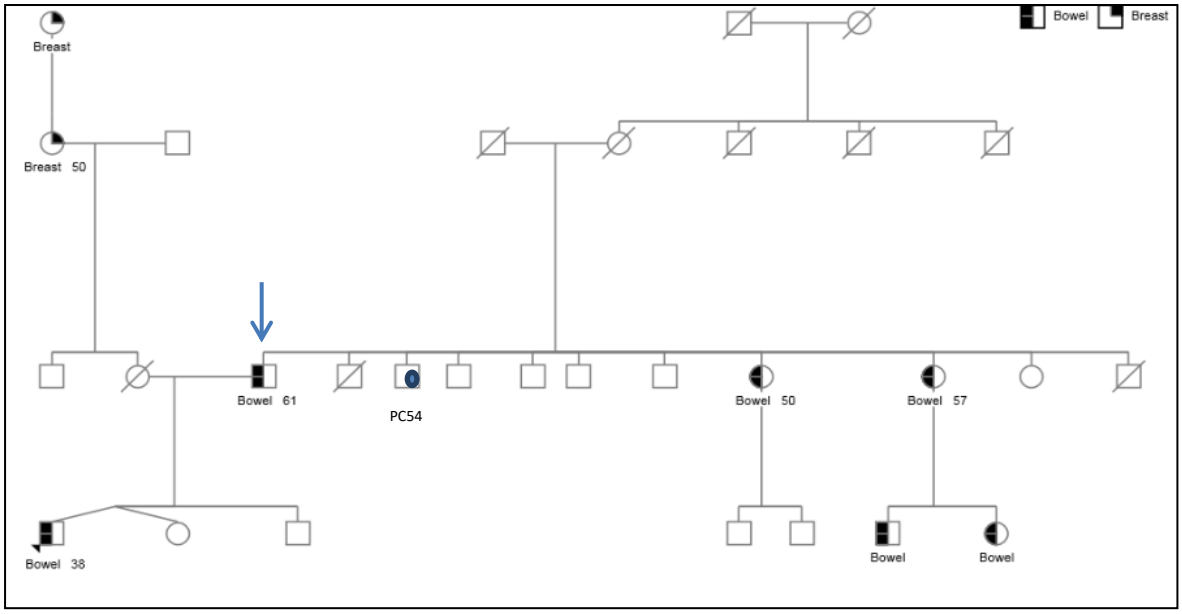



**Figure 5-1. *TMCC3* variant carrying family. Pedigree of paternal uncle (PC54) with colonic adenomatous polyposis and nephew (PC41, proband, shown with an arrowhead) with young-onset CRC.**

The affected family member who agreed to participate in this study was the father of PC41 and brother of PC54 (Figure 5.1), hence he would be an obligate carrier of the novel *TMCC3* missense variant shared between PC54 and PC41. Interestingly, PC54 had multiple metachronous colonic polyps, while PC41 and his father had CRC without polyps identified. The NM_020698.2 (*TMCC3*):c.362A>G variant that was identified in both PC41 and PC54 through intra-familial analysis, was also present in the father of the proband. There were no *TMCC3* LoF variants or novel missense variants in the validation cohort, hence there is no supporting evidence from the validation cohort that *TMCC3* mutations predispose to CRC. The other affected individuals (two paternal aunties with CRC aged in their 50s and 2 paternal cousins with CRC aged in their 30s) who would have been more informative were invited to participate in the study by their FCC but declined.

## 5.4. Discussion

In this chapter, 50 candidate CRC-predisposing genes were re-sequenced in a validation cohort of 333 individuals with young onset and/or familial CRC. Very few candidate CRC-predisposing genes had LoF variants in the validation cohort. Possible explanations are that familial/young-onset CRC is genetically heterogeneous and that mutations in these genes are a very rare cause of CRC, or that the remaining genes do not predispose to CRC, or a proportion of these familial/young-onset CRC cases do not have a genetic aetiology.

### 5.4.1. Evaluation of candidate CRC-predisposing genes in the validation cohort

Fourteen candidate genes were found to have a LoF variant in at least one individual in the validation cohort, with seven of these genes having LoF variants in 2 or more individuals. Of the 7 genes with a LoF variant in multiple individuals, 5 (*WNT10A, NEDD4, MS4A15, TNFRSF10B and LRRK2*) appear to have a higher frequency of LoF variants in the validation cohort than would be expected, using the non-TCGA subset of the ExAC dataset to estimate the frequency of LoF variants in the population.

*WNT10A* encodes a ligand involved in the non-canonical Wnt signalling pathway (Fan et al., 2017). *WNT10A* antagonises WNT3A-induced β-catenin/TCF activity in the liver, and inhibits cell proliferation and promotes liver differentiation of implanted liver cell progenitors (Fan et al., 2017). Homozygous mutations in *WNT10A* result in a form of ectodermal dysplasia (Adaimy et al., 2007); CRC has not been reported in these cases.

*NEDD4* is a homologous to E6AP C-terminus (HECT) E3 ubiquitin ligase with conflicting findings in CRC in the literature. Recurrent insertions in an insertional mutagenesis study in mice (March et al., 2011), loss-of-heterozysity in CRC and CRC cell lines (Martin et al., 2007) and overexpression in CRC cell lines (Eide et al., 2013) have been reported. This may reflect either heterogeneity in disease biology, or stochastic events.

The function of the gene product of *MS4A15,* a membrane-spanning protein, is unknown.

*TNFRSF10B* is a TNF-superfamily receptor that induces apoptosis and activates NF-kappa-B when bound to TRAIL (Sheridan et al., 1997). A different truncating mutation to the ones identified in the validation cohort has been reported in a head and neck tumour and germline of a patient (Pai et al., 1998).

*LRRK2* is a protein kinase that has been postulated to carry driver mutations in cancer (Greenman et al., 2007). Somatic mutations and downregulation of LRRK2 are found in squamous cell carcinomas of the lung (Meng et al., 2019) Germline point mutations in *LRRK2* are associated with autosomal dominant Parkinson's Disease (Paisán-Ruíz et al., 2004, Zimprich et al., 2004)

*GPR144* had low sequencing coverage in the non-TCGA subset of ExAC, hence the apparent increased frequency of Lof variants in the validation cohort may be due to under-detection in the non-TCGA subset of ExAC. One of the 2 LoF variants in *GPR144* occurred in the last exon, hence there is no strong evidence for CRC-predisposition.

Of the 7 genes with 1 individual with a LoF variant in the validation cohort, 5 genes (*GALNT8, ASCL3, ERCC3, NOTCH4* and *FKBP10*) had approximately the same frequency of LoF variants in the non-TCGA subset of ExAC, and hence are unlikely to be CRC-predisposing genes. Of the remaining genes, *INPP5E* had a LoF variant in the last exon, hence the functional significance of this LoF variant is unclear. There was however an individual with a LoF variant that was not in the last exon in the discovery cohort, in addition to the shared novel missense variant in the intra-family analysis used to prioritise *INPP5E* as a candidate CRC-predisposing gene. Only 6 LoF variants were identified in 53 105 individuals in the non-TCGA subset of ExAC. Knockdown of *INPP5E*, a spindle assembly checkpoint phosphatase, results in chromosomal instability in human fibroblasts (Sierra Potchanant et al., 2017a).

Evaluation of *WNT10A, NEDD4, MS4A15, TNFRSF10B, LRRK2, INPP5E, UFL1* in a larger cohort of individuals with young-onset/familial CRC would be of interest. A control population was not sequenced using the Haloplex Enrichment System kit. This would have been useful for controlling for technical factors such as gene coverage, sequencing or alignment errors. However, it would also have added to the cost of the project and substantially reduced the number of validation cases that could be sequenced. While a different sequencing methodology and platforms were used for the non-TCGA subset of the ExAC dataset (WES across different projects that make up ExAC), this large control dataset is still useful for excluding genes if the frequency of LoF variants in a gene is equivalent or greater than in the validation cohort, as a LoF variant in these genes is then unlikely to predispose to CRC.

No rare LoF variants were seen in the validation cohort for *BUB1B, BUB1, BUB3, NTHL1*, or *KMT2C*, the 5 candidate CRC-predisposing genes identified from the literature. *BUB1B* and *KMT2C* were originally identified in a single individual and 4 related individuals from a single family respectively, hence these genes have not been identified as recurrently mutated in a validation cohort. However, *BUB1*, *BUB3* and *NTHL1* variants were identified in unrelated individuals with young onset/familial CRC (*BUB1, BUB3*) or colonic adenomas (*NTHL1*) in the literature. As our validation cohort was not predominantly comprised of individuals with colonic adenomas, the absence of biallelic *NTHL1* variants may be due to such variants only giving rise to this specific phenotype.

Targeted sequencing of 192 individuals with CRC developing at age 50 years or younger did not identify compound heterozygous or homozygous mutations in *BUB1B* (Hahn et al., 2016). However neither this paper nor our data specifically looked at intronic sequences or large deletions (as in the original report), nor were cases select based on mosaic variegated aneuploidy. Hence, while *BUB1B* homozygous or compound heterozygous mutations are not common in young onset/familial CRC, individuals with mosaic variegated aneuploidy and CRC may be considered for further testing of the *BUB1B* gene.

Only missense variants were identified in *BUB3* in the original study (de Voer et al., 2013). One rare missense variant in BUB3 was identified in our validation cohort, however this variant was predicted to be benign by PolyPhen and tolerated by SIFT.

### 5.4.2. Analysis for LOH of the PPV in the corresponding CRC

The PPVs could not be assessed in some of the tumour samples despite multiple attempts, possibly due to the age of the tumour blocks (up to 18 years from fixation) or the regions being amplified. As a result, information on LOH was not available for the majority of the PPVs to either support or refute their role in CRC predisposition.

### 5.4.3. Challenges of performing segregation analysis for the PPV in *TMCC3*

The segregation analysis identified the PPV NM_020698.2 (*TMCC3*):c.362A>G in an additional affected family member, however as this individual was an obligate carrier this was not informative. The function of *TMCC3* is unknown. It would have been informative to know if this variant was also present in the other affected members of the family, as there is a 25% probability of the father of the proband sharing a variant by chance, being a FDR of the other two individuals. Segregation studies usually involve testing affected and unaffected members of the family to determine if a variant segregates with disease. In the absence of being able to test the other more distally related affected family members in PC41's family, recruitment of unaffected family members was not performed.

### 5.4.4. Conclusion

Despite the candidate genes having multiple LoF variants or other supporting evidence from the discovery cohort for a role in CRC predisposition, the vast majority did not have any LoF variants in the validation cohort emphasising the highly heterogeneous causes of the remaining CRC families and early onset cases. This heterogeneity is consistent with the findings of a large case control study of early onset familial CRC cases that did not identify novel CRC-predisposing genes accounting for >1% of familial CRC risk (Chubb et al., 2016). *WNT10A, NEDD4, MS4A15, TNFRSF10B, LRRK2, INPP5E and UFL1*, should be evaluated in larger case control datasets to determine if there is a significant increase in PPV in young onset/familial CRC cases compared with unaffected controls, as a chance occurrence in the validation dataset cannot be excluded. No LoF variants were found in *NTHL1, KMT2C, BUB1B, BUB1* or *BUB3,* indicating that these genes are not common CRC-predisposing genes in this Australian cohort.

# Chapter 6. Summary and Future Directions

## 6.1. Overview

There is a clinical need to better understand the genetic predisposition to developing CRC in individuals with familial/young onset CRC that are referred to FCCs, in whom an inherited pathogenic variant in a known CRC-predisposing gene is not identified. Identification of novel CRC-predisposing genes in such individuals, if validated by a clear increase in occurrence of pathogenic variants in these genes in CRC cases compared to controls, together with segregation of the variant with disease within affected families, would enable identification and exclusion of family members who do not carry the causative variant from intensified cancer surveillance or prophylactic surgery. Follow-up studies of individuals who carried causative variants in novel CRC-predisposing genes would then be carried out to define the magnitude of increased CRC risk associated with novel genes, in order to inform risk reduction strategies.

When this study began, the majority of FCCs were still using a sequential single-gene sequencing approach to diagnose a germline predisposition to CRC, and only known CRC-predisposing genes were analysed. As these individuals have an increased a priori risk of a germline predisposition to CRC, it was postulated that a mutation in a novel gene may be responsible for the familial CRC or young age of onset of CRC seen in these individuals.

This thesis describes the use of WES of germline DNA to investigate a cohort of 54 individuals (from 49 families), recruited from Australian FCCs or a prospective CRC cohort in whom standard clinical genetic testing had not identified a pathogenic variant in a known CRC-predisposing gene. Three unrelated individuals were subsequently found to have a mutation in a known CRC-predisposing gene through WES and were excluded from the discovery cohort. The final discovery cohort of 51 unexplained individuals was analysed using inter-family, intra-family, CRC pathway and gene constraint approaches to prioritise 45 candidate CRC-predisposing genes.

Analysis of CRCs for LOH of PPVs in tumour tissue was carried out where possible, but this was limited by lack of availability of tumours, low tumour purity and possibly the age of the tumour blocks. Of the 4 tumours that were successfully sequenced, no LOH of the PPV was found for any of the corresponding novel candidate genes: *FKBP10* in 2 samples, *ASCL3* and *JAG2* in one sample each. Affected family members were invited to participate but did not respond, hence segregation data could not be obtained for the PPV in the candidate CRC-predisposing genes. Targeted gene re-sequencing of these candidate CRC-predisposing genes, plus 5 candidate CRC-predisposing genes reported in the literature, was performed in a validation cohort of 333 individuals with familial or young-onset CRC. Seven candidate CRC-predisposing genes (*WNT10A, NEDD4, MS4A15, TNFRSF10B, LRRK2, INPP5E* and *UFL1*) had PPVs identified in the validation cohort, at a rate higher than expected in the non-cancer population, when estimated using the non-TCGA subset of the ExAC dataset.

## 6.2. Establishing a Discovery Cohort for Novel CRC-Predisposing Genes

The majority of individuals in the discovery cohort represent patients seen in FCCs who are suspected of having a genetic predisposition to CRC, but in whom a mutation in a known CRC-predisposing gene was not identified by clinical testing. Hence this study addressed a clinical need in the FCCs by offering WES to such individuals with the aim of identifying novel CRC-predisposing genes. Other studies using WES of germline DNA in patients with familial or young-onset CRC have identified patients from a national CRC biorepository (Chubb et al., 2016) or registry (Derycke et al., 2013), research consortium (Esteban-Jurado et al., 2014), hospital CRC clinics (Zhang et al., 2015) or CRC clinical trials (Smith et al., 2013). FCCs as a source of recruitment are advantageous due to verification of a family history of cancer for each individual as part of their work-up. This verification is usually not available through other sources of recruitment. Conversely, recruitment from national biorepositories/registries/ consortia or clinical trials allows ready access to a large number of individuals with CRC, whereas recruitment from FCCs or clinics depends on the rate of referral of individuals with familial CRC.

The discovery cohort was clinically heterogeneous, reflecting the heterogeneity that is seen in FCCs in the patient cohort in which no known CRC-predisposing gene mutation is found. This included patients with CRC without polyposis, patients with polyposis with or without CRC, and patients with different polyp types. It is likely that the cohorts in biorepositories, CRC clinics and clinical trials are similarly heterogeneous, with varying degrees of annotation with regards to colonic polyp subtypes.

## 6.3. Analysis Methods for Identifying Novel Candidate CRC-Predisposing Genes

### 6.3.1. The use of WES to identify novel cancer-predisposing genes in familial cancer

Following on from the first successful applications of germline WES in identifying novel predisposing genes in monogenic diseases (Ng et al., 2010b, Ng et al., 2010a, Hoischen et al., 2010), WES has been used to study the genetic basis of complex diseases including familial cancer, with the aim of identifying rare causative variants in known or novel cancer-predisposing genes. While there have been some successes in familial cancers, mostly involving very rare cancer subtypes with specific phenotypes (Jones et al., 2009b, Popova et al., 2013, Witkowski et al., 2014, Gara et al., 2015), the use of WES or whole genome sequencing to identify novel cancer-predisposing genes in more common familial cancers has been hampered by genetic heterogeneity (Chubb et al., 2016, Roberts et al., 2016). A very specific subtype of a rare cancer, with an associated rare phenotype, is more likely to reflect an underlying genotype-phenotype correlation akin to a monogenic disease, whereas locus heterogeneity giving rise to the more common cancers such as CRC reduces the power of WES to identify recurrently mutated genes in familial cancer cohorts. WES also does not comprehensively assess CRC risk alleles identified from GWAS studies, the majority of which are in non-coding regions of the genome. Polygenic risk conferred by these alleles may account for a proportion of young onset or familial CRC cases.

### 6.3.2. Analytical approaches to PPV identified by WES in order to prioritise candidate CRC-predisposing genes

#### *6.3.2.1. Inter-family analysis and intra-family analysis*

The inter-family and intra-family approaches are most commonly used in the literature to identify candidate cancer-predisposing genes (Derycke et al., 2013, Palles et al., 2013, Gylfe et al., 2013a, Esteban-Jurado et al., 2014, Nieminen et al., 2014, Tanskanen et al., 2015). The lack of overlap in the candidate gene lists identified in these studies suggests that familial CRC is a genetically heterogeneous disease, or that monogenic inheritance is not the basis of familial CRC in at least some of the individuals recruited into WES studies. In intra-family analysis, when a monogenic cause is suspected, phenocopies or the co-occurrence of different monogenic causes of CRC within the family can confound analysis. Intra-family analysis is also dependent on having multiple affected family members, preferably in multiple generations, in order to reduce the overlap of variants that are shared between them, so as to increase the chance of differentiating a PPV from a variant that is shared by chance. Families with multiple affected individuals also facilitates segregation analysis of a PPV within the family to support the association between the variant and disease. The identification of *POLE* and *POLD1* is a successful application of an intra-family analysis, however the requirement of 3805 highly selected young-onset CRC cases with multiple adenomas to detect a further 12 individuals with the POLE p.L424V pathogenic variant demonstrates the potential difficulty of validating

candidate CRC genes without a specific accompanying clinical phenotype (multiple adenomas) to reduce clinical heterogeneity, which may also then reduce genetic heterogeneity.

#### 6.3.2.1.1. Intra-family analysis in a rare clinical phenotype

This study included the intra-family analysis of a father and daughter with AMTs, exceedingly rare tumours with an incidence of less than 1 in a million individuals (Shaib et al., 2017). Only 2 affected families (sibling pairs in both cases) have previously been reported in the literature (Racek et al., 2011, Shih et al., 2001). Limited genetic analysis (LOH of *APC* in one family, and sequencing of the MMR genes in one individual of the other family) had been performed.

Through WES, inherited pathogenic variants in the known CRC-predisposing genes *APC, MUTYH, MLH1, MSH2, MSH6, PMS2, BMPR1A, SMAD4, POLE, POLD1* were excluded in the father and daughter with AMTs, providing the first in-depth analysis of known CRC-predisposing genes in FDRs with AMTs. *STK11* mutations could not be confidently excluded due to low coverage. Seventeen private PPVs were shared between father and daughter. Of these 17 variants, 4 variants were within regions of LOH identified in PMP from the father. When tumour DNA from the father's AMT was used to confirm LOH identified in PMP (that was used to obtain adequate DNA for genome-wide copy number analysis), there was no loss of the wildtype allele within regions of LOH. One variant had loss of the mutant allele in a region of LOH, and was excluded as it is unlikely to predispose to AMTs. Comprehensive assessment of hypocellular tumours for second hits is challenging due to the low amount of DNA obtained from each sample. In order to obtain enough DNA for copy number analysis, the peritoneal PMP samples were analysed. It is important to confirm LOH in the primary tumour as regions of LOH in metastases may not be present in the primary tumour, as was seen in this case. However, as LOH is not the only mechanism for a second hit, the sixteen candidate genes identified from this AMT family warrant further study in other familial AMT cases. It is possible that this family has an inherited genetic predisposing factor that was not captured by WES, such as a variant in a gene regulatory region or a structural variant. A shared environmental factor, or chance occurrence of AMTs in FDRs is possible although much less likely.

### *6.3.2.2. CRC pathway analysis and gene constraint analysis*

Pathway analysis is used in a smaller number of studies, as this is a candidate gene approach that is not unbiased. The 3 known CRC-predisposing genes containing causative variants identified in the discovery cohort would only have been identified by pathway analysis, with the exception of *APC* that would also have been prioritised due to its high gene constraint score. The gene constraint approach was based on the hypothesis that CRC-predisposing gene mutations would be selected against in the population, and leverages the constraint scores obtained from the large ExAC dataset (Lek et al., 2016a). Lek *et al* proposed that genes that showed constraint may be disease-causing, hence we looked for genes under constraint in our validation dataset. It would have been interesting to re-sequence all 30 constrained genes with a single LoF variant in the discovery cohort, however due to cost constraints only 3 with plausible gene function were re-sequenced in the validation cohort. LoF-intolerant genes did not have high quality LoF variants in 2 or more families in the discovery cohort, suggesting that these genes are not commonly mutated in familial CRC.

### 6.3.3. Prioritisation of LoF variants in the 4 analytical analyses

The 4 analytical approaches prioritised 45 candidate CRC-predisposing genes for re-sequencing in a validation cohort. All 4 analyses focused on LoF variants, apart from the intra-family analysis that leveraged CADD as an in-silico score that incorporates 63 different annotations to rank deleteriousness of a base substitution, together with the knowledge that the variant was inherited in both affected individuals within a family. Classifying missense variants is challenging even in known CRC-predisposing genes, and most WES studies of familial CRC have focused on LoF variants for this reason.

## 6.4. Re-sequencing of candidate CRC-predisposing genes in a validation cohort

A targeted gene panel was designed consisting of the 45 candidate CRC-predisposing genes identified from the discovery cohort and 5 candidate CRC-predisposing genes identified from the literature. Re-sequencing of these 50 candidate CRC-predisposing genes in a validation cohort of 333 young-onset/familial CRC cases, prioritised 10 genes (*WNT10A, NEDD4, MS4A15, TNFRSF10B, LRRK2, INPP5E, UFL1, GOLGA5, HTR5A and FKBP10*) with a PPV variant in at least one individual. The majority of the candidate genes did not have a PPV in the validation cohort, implying that they are not genuine CRC-predisposing genes, or that these genes are extremely rare causes of CRC when mutated.

### 6.4.1. Genetic heterogeneity of familial CRC affects the required size of the validation cohort

A study by Chubb *et al.* used WES of 1061 familial CRC cases and 1644 individuals without a history of cancer, to identify CRC-predisposing genes with an enrichment of nonsense and frameshift variants in cases (Chubb et al., 2016). In this study, the familial CRC cases were diagnosed at or before age 55 years, with at least one FDR with CRC. The control set comprised 1644 individuals born in 1958 without a history of cancer. The cohorts were European and UK population respectively, comparable to the majority of the Australian population. Nonsense and frameshift variants with a minor allele frequency <1% were summed for each gene. While a statistically significant increase in *APC, MLH1* and *MSH2* nonsense or frameshift variants were seen in cases, it is somewhat sobering that the analysis of the known CRC-predisposing genes did not identify a statistically significant enrichment of frameshift or nonsense variants in the familial CRC cohort for known CRC-predisposing genes including *MSH6, PMS2, STK11, SMAD4* and *BMPR1A.* Chubb *et al* then restricted their case control analysis to 863 cases without a pathogenic variant in a known CRC-predisposing gene in order to identify novel CRC-predisposing genes. When compared to the 1644 controls, no novel genes showed a significant increase in nonsense or frameshift variants in cases at an exome-wide significance ($p=8.0X10^{-7}$ using Bonferroni correction).

Taking the missed known CRC-predisposing genes as a reference point, while *STK11, SMAD4* and *BMPR1A* are associated with a specific phenotype and onset in childhood that may not have been included in the cases, *MSH6* and *PMS2* mutation carriers would reasonably be expected to be included in the cases. These 2 genes have reduced penetrance and account for <1% of CRC. The fact that known CRC-predisposing genes were not identified despite the use of over 1000 CRC cases and over 1000 controls illustrates the challenge of relying on statistical significance alone to identify novel rare CRC-predisposing genes that account for <1% of CRC.

Hence, due to the genetic heterogeneity of familial CRC, the candidate CRC-predisposing genes with LoF variants in the validation cohort of this thesis may be CRC-predisposing, however further evidence is required from segregation and/or functional studies in additional cases.

### 6.4.2. Assessment of LoF variants in candidate genes proposed in the literature

Of the 5 candidate CRC-predisposing genes identified from the literature (*NTHL1*, *BUB1, BUB1B, BUB3* or *KMT2C)* that were included in the targeted gene panel, no LoF mutations were found in the validation cohort. Broderick *et al* compared the incidence of pathogenic variants in proposed CRC-predisposing genes *NTHL1, BUB1, BUB3, RPS20, FANCM, FAN1, TP53, LRP6* and *PTPN12* in WES data from 863 individuals with familial CRC (diagnosed at or below the age of 55 years), compared to WES data from a control set of 1604 individuals without CRC (Broderick et al., 2016). This dataset was also used in the Chubb *et al.* study. No LoF variants were found in *BUB1* or *BUB3* in cases, or controls. No homozygous LoF variants in *NTHL1* were found in cases or controls. A 41-year-old case with CRC and polyposis was found to have compound heterozygous truncating mutations (NTHL1 p.Q90*/p.Q287*), while no controls harboured compound heterozygous mutations.

In this systematic review, Broderick et. al concluded that there was support for *NTHL1* being a CRC-predisposing gene due to characteristic C:G>T:A transitions seen in associated tumours, together with multiple autosomal recessive cases reported in the literature associating *NTHL1* with CRC. The first Australian case report of homozygous *NTHL1* nonsense mutations in a patient with adenomatous polyposis and bladder cancer was recently reported (Groves et al., 2019).

The data presented in this thesis, and that of Broderick *et al* suggest that *NTHL1* mutations are not a common cause of young-onset/familial CRC. It appears that *NTHL1* mutations are more likely to be identified in colonic polyposis patients (Belhadj et al., 2019), particularly if they have also developed another cancer associated with *NTHL1* deficiency (Grolleman et al., 2019). While the polyp status of the validation cohort in this thesis is not known, the discovery cohort included 9 individuals with 10 or more unexplained colonic polyps, and no *NTHL1* LoF mutations were seen in both cohorts. None of the 9 individuals had a second cancer associated with *NTHL1* deficiency.

For the remainder of genes tested by Broderick et al, one LoF variant was found for *RPS20* with none seen in controls, while no evidence of an excess of LoF variants in cases was seen for the other genes. *RPS20* was not included in the targeted gene panel as it had only been described in one CRC family.

## 6.5. Limitations

This study used WES to identify PPV predisposing to familial CRC. While WES is used with the intent of assessing the exome, interrogation of the entire exome is not fully achieved in practice. Protein-coding exons in the human genome that have not been annotated will be missed as these are not included in probe design for exome enrichment. Some protein-coding regions are difficult to target with probes (Majewski et al., 2011). Sequencing coverage is not uniform across targeted regions, and can be affected by G-C content (Bentley et al., 2008). Large structural variants are not detected by WES. Variant annotation for WES is performed for all known gene transcripts. A variant that is called in WES may have different predicted functional consequences depending on its transcript context. Knowing which transcript to select for variant analysis can be challenging as transcript expression in the colon is not known for every gene.

As WES does not target non-coding regions of the human genome apart from splice sites, structural variants and SNVs in these regions will be missed. As previously alluded to, CRC risk alleles identified from GWAS studies are not routinely captured through WES. Whole genome sequencing would cover the regions missed by WES, however as most causative variants for Mendelian disease were identified in the exome, and because the amount of sequencing required for whole genome sequencing made the cost prohibitive in 2013 when this study commenced, WES was chosen for this study.

Of the variants identified through WES, missense variants were only assessed in the intra-family analysis, and other variant types such as synonymous variants and splice site variants (apart from essential splice site variants) were not considered in the analyses, due to the uncertainty around the pathogenicity of such variants. It is possible that the causative variants predisposing to CRC in some of the families in this study were within these variant types. The validation cohort size was relatively small, hence a chance occurrence of a PPV in the candidate CRC-predisposing genes could not be excluded. There was marked clinical heterogeneity in the discovery cohort, hence the absence of a PPV in the validation cohort may be due to the latter not including individuals with the same specific clinical phenotype.

## 6.6. Future Directions

While WES has been successfully applied to monogenic diseases, it is becoming apparent that its application in complex disease is hampered by disease heterogeneity and a resultant decrease in the ability to discriminate disease-predisposing genes solely by statistical comparison of cases and controls, due to the need to correct for multiple testing.

Recent success in identifying CRC-predisposing genes has come from intra-family analysis (Palles et al., 2013), and inter-family analysis in a country (the Netherlands) with a fairly homogenous population (Weren et al., 2015). Both these scenarios increase the chance of identifying a monogenic cause of CRC. In Australia, where the population is more heterogeneous, focusing on intra-family analysis of large families with multiple affected individuals may be preferable to inter-family analysis.

Concurrent WES of corresponding CRCs may identify genomic clues such as particular mutation signatures (Alexandrov et al., 2013) that may be helpful in subgrouping familial CRC, or may possibly point to an affected pathway in some instances, in a similar way to how *MUTYH* was identified from the particular transversions seen in *APC* in corresponding CRCs. Genomic characterisation of tumours has advanced considerably since the commencement of this study, for example with mutation signatures and tumour mutational burden defining a subset of CRCs arising from MMR protein deficiency. Both tumour mutation signatures and tumour mutational burden correlate with mismatch repair deficiency (Stadler et al., 2016), while CRCs with *POLE* or *POLD1* mutations have been found to be ultramutated (The Cancer Genome Atlas, 2012, Palles et al., 2013).

Due to the genetic heterogeneity of familial CRC, a highly selected validation cohort with an associated phenotype (if one is identified) may be required to demonstrate enrichment in cases compared to controls. For example, a validation cohort of 3805 familial/young-onset cases of CRC that included families with colorectal adenomas was required to identify 12 individuals with the pathogenic *POLE* p.Leu424Val (NM_006231; c.1270C>G) variant. Sizable cohorts of rare phenotypes of CRC are harder to access, however such carefully phenotyped cohorts may be helpful to decrease the genetic heterogeneity within the validation cohort, provided a genotype-phenotype correlation can be identified. Identification and collation of carefully phenotyped subgroups of CRC will firstly require pathologists to keep abreast of updates in CRC sub-classification and adhere to such guidelines when reporting, in order to facilitate targeted recruitment for future studies. Achieving an adequate sample size of particular subgroups of CRC will likely require national and international research collaborations. A specific phenotype may however still have considerable locus heterogeneity, for example the adenomatous polyposis phenotype is not specific to mutations in a single gene.

Whole genome sequencing will interrogate the 99% of the genome not covered by WES. Copy number variation, which is known to predispose to hereditary CRC (eg. GREM1 duplication), will be assessed by whole genome sequencing. The difficulty lies in interpretation of variants in so-called non-coding regions, and although the Encyclopedia of DNA Elements project (The

Encode Project Consortium et al., 2012) has begun to shed light on the functional elements within the entire genome, much remains to be discovered. Whole genome sequencing will however allow assessment of the contribution of polygenic risk, but may require very large cohorts if the effect size is modest.

## 6.7 Final Conclusions

The aim of this thesis was to perform WES of germline DNA on an Australian population deemed to be at increased risk of harbouring a novel inherited pathogenic variant predisposing to CRC, to identify candidate CRC-predisposing genes. Following the identification of 45 candidate CRC-predisposing genes from a discovery cohort of 51 individuals with familial/young-onset CRC, re-sequencing in a validation cohort of 333 cases prioritised 7 genes (*WNT10A, NEDD4, MS4A15, TNFRSF10B, LRRK2, INPP5E, UFL1*) with a LoF PPV variant in at least one individual. Due to the small numbers of individuals in the validation cohort with a PPV, a chance occurrence of PPVs in these seven genes cannot be excluded. Re-sequencing of these 7 genes in a larger cohort of familial CRC cases, together with segregation studies and studies of the functional effect of the corresponding gene mutation in tumour, may help to determine if these genes are CRC-predisposing genes when mutated. Re-sequencing of the 16 genes with a PPV shared by a father and daughter with AMTs, in other families with AMT, together with comprehensive assessment for a second hit in tumour tissue, may help determine if one of these novel genes predisposes to the development of AMT when mutated. No *NTHL1*, *BUB1, BUB1B, BUB3* or *KMT2C* LoF variants were seen in the validation cohort, suggesting that LoF variants in these genes are not a common cause of familial CRC in an Australian validation cohort.

**Appendix 1. Supplementary Material for 'Candidate genes predisposing to familial appendiceal mucinous tumours presenting with pseudomyxoma peritonei'**

**Supplementary Data 1. Ranking of gene transcript according to predicted functional consequence of a variant**

| Functional Consequence | Rank |
|---|---|
| Nonsense, frameshift | 1 |
| Essential Splice Site | 2 |
| Missense | 3 |
| Other | 4 |

**Supplementary Data 2. PCR primers for validation of loss-of-function (LoF) variants and missense variants with Combined Annotation Dependent Depletion scaled score ≥10**

| Gene | Sequencing direction | Primer Sequence | PCR Product Size |
|---|---|---|---|
| *RHBDL2* | Forward | CTGAGAAGAGGGAGGAAGCC | 244 |
| *RHBDL2* | Reverse | cctcagcctccatgactaca | 244 |
| *FGFR4* | Forward | TGGTCTTTTGGGATCCTGCT | 185 |
| *FGFR4* | Reverse | ggaggaggaggactggaaag | 185 |
| *CNTN2* | Forward | cttggcaaaggttgggtgg | 203 |
| *CNTN2* | Reverse | GACAAAGTACTGGGCATCGG | 203 |
| *RANBP2* | Forward | AAGAATGGAAAGAACGTGGGA | 220 |
| *RANBP2* | Reverse | GCAAGTTGTTCTGGTTTTGGC | 220 |
| *HSPB6* | Forward | GGAGGAGCAGGATGGAGATC | 242 |
| *HSPB6* | Reverse | gtggtcgagttcacccctc | 242 |
| *ZNF747* | Forward | GAGACCTACGGCCACCTG | 225 |
| *ZNF747* | Reverse | ACAGTTCGGCCTTCTCCTC | 225 |
| *BRINP3* | Forward | tgttcttcaagGGCTTCAAGT | 150 |
| *BRINP3* | Reverse | TCTGGAAATTTGGGACCACAG | 150 |
| *ASIC1* | Forward | GATGGCAGATGAAAAGCAGC | 244 |
| *ASIC1* | Reverse | gcctgctccttgtctcctc | 244 |
| *TNFRSF1B* | Forward | CTTTCGGTCACAGCTGGAGA | 226 |
| *TNFRSF1B* | Reverse | GAACTTGGCCCAGAAAGAGC | 226 |
| *RANBP6* | Forward | ACCTTCCTCTTAGATGCCGT | 236 |
| *RANBP6* | Reverse | ATCAAATTCCTGGCCAGCAC | 236 |
| *LSR* | Forward | ccatgcactagggcttcag | 237 |
| *LSR* | Reverse | ctcccgagaagcccaactc | 237 |
| *MTERFD3* | Forward | TGACAGCTGCACCTAATGTT | 206 |
| *MTERFD3* | Reverse | ACCTTGCTCCTGGAGAAATTC | 206 |
| *PAX1* | Forward | gtagtgatccgacgcctctg | 239 |
| *PAX1* | Reverse | CTCCGCGAGTCCTCCGATC | 239 |
| *EGFR* | Forward | tctctttcacttcctacagATGC | 150 |
| *EGFR* | Reverse | aggacagtcagaaatgcagga | 150 |
| *THOP1* | Forward | CTACGACGCCCAGTACTACG | 300 |
| *THOP1* | Reverse | caggccaaaggtgtgcttt | 300 |

**Supplementary Data 3. M13 PCR primers for determination of loss-of-heterozygosity (LoH) in tumour tissue**

| Gene | Sequencing direction | Primer Sequence | PCR Product Size |
|---|---|---|---|
| *REEP5** | Forward | GTAAAACGACGGCCAGTttgtggtccccagGTGTC | 108 |
| *REEP5** | Reverse | AACAGCTATGACCATGcGAGATGTAGGCTGGGTAGC | 108 |
| *EXOG** | Forward | GTAAAACGACGGCCAGTaagagccactaattgtgtgtgtg | 136 |
| *EXOG** | Reverse | AACAGCTATGACCATGACACCTTGCCTCTGTTCCAG | 136 |
| *RANBP2* | Forward | GTAAAACGACGGCCAGTATGGAAAGAACGTGGGATTG | 108 |
| *RANBP2* | Reverse | AACAGCTATGACCATGTTGCACAGATTTTCAATACTTGCT | 108 |
| *RANBP6* | Forward | GTAAAACGACGGCCAGTACAAATGGCTGCCGCACT | 113 |
| *RANBP6* | Reverse | AACAGCTATGACCATGCAGCCAGAATCAGTTCAATCTT | 113 |
| *TNFRSF1B* | Forward | GTAAAACGACGGCCAGTAATGTGCCTTTCGGTCACAG | 110 |
| *TNFRSF1B* | Reverse | AACAGCTATGACCATGTGGTTAACTGGGCTTCATCC | 110 |
| M13 Forward Primer | | GTAAAACGACGGCCAGT | N/A |
| M13 Reverse Primer | | AACAGCTATGACCATG | N/A |

*Primers also used for validation of variants in germline DNA

**Supplementary Data 4. Primers for *REEP5* exons**

| Exon | Sequencing direction | Primer Sequence | PCR Product Size |
|---|---|---|---|
| Exon 1 | Forward | AGTCGCCGCTCCAGTCTAT | 266 |
| Exon 1 | Reverse | gagtcctctccctgcttcct | 266 |
| Exon 2 | Forward | cgagtcgagaggaaaactcc | 292 |
| Exon 2 | Reverse | aattgcggatgtgcagtcta | 292 |
| Exon 3 | Forward | gaaagtgtatttgaaaatctttcattg | 298 |
| Exon 3 | Reverse | gccacagatgcacaacattc | 298 |
| Exon 4 | Forward | cccatccccagagtagtgaa | 364 |
| Exon 4 | Reverse | gctgaggggtggtgataaga | 364 |
| Exon 5 | Forward | atgatcgcacctgaacatga | 287 |
| Exon 5 | Reverse | TGTTTCCAAGGCAACATTATT | 287 |

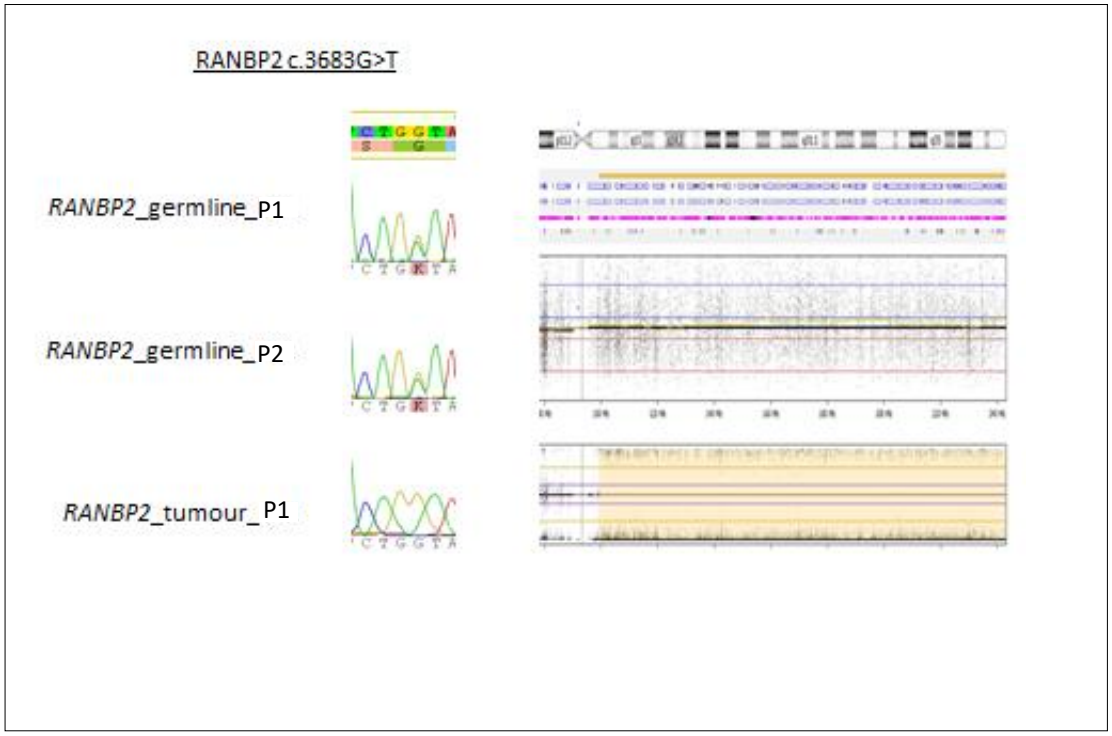
**Supplementary Data 5. Coverage of familial colorectal cancer genes in whole exome sequencing**

| Sample | Gene | Mean coverage |
|---|---|---|
| **P1** | *APC* | 74.59 |
| **P2** | *APC* | 73.25 |
| **P1** | *BMPR1A* | 154.82 |
| **P2** | *BMPR1A* | 201.13 |
| **P1** | *EPCAM* | 80.65 |
| **P2** | *EPCAM* | 93.60 |
| **P1** | *GREM1* | 68.52 |
| **P2** | *GREM1* | 92.03 |
| **P1** | *MLH1* | 158.03 |
| **P2** | *MLH1* | 138.76 |
| **P1** | *MSH2* | 108.17 |
| **P2** | *MSH2* | 105.14 |
| **P1** | *MSH6* | 76.81 |
| **P2** | *MSH6* | 75.05 |
| **P1** | *MUTYH* | 125.69 |
| **P2** | *MUTYH* | 88.85 |
| **P1** | *NTHL1* | 172.70 |
| **P2** | *NTHL1* | 127.78 |
| **P1** | *PMS2* | 129.68 |
| **P2** | *PMS2* | 130.68 |
| **P1** | *POLD1* | 96.10 |
| **P2** | *POLD1* | 85.39 |
| **P1** | *POLE* | 193.59 |
| **P2** | *POLE* | 186.60 |
| **P1** | *SMAD4* | 205.47 |
| **P2** | *SMAD4* | 240.61 |
| **P1** | *STK11* | 3.46 |
| **P2** | *STK11* | 2.08 |

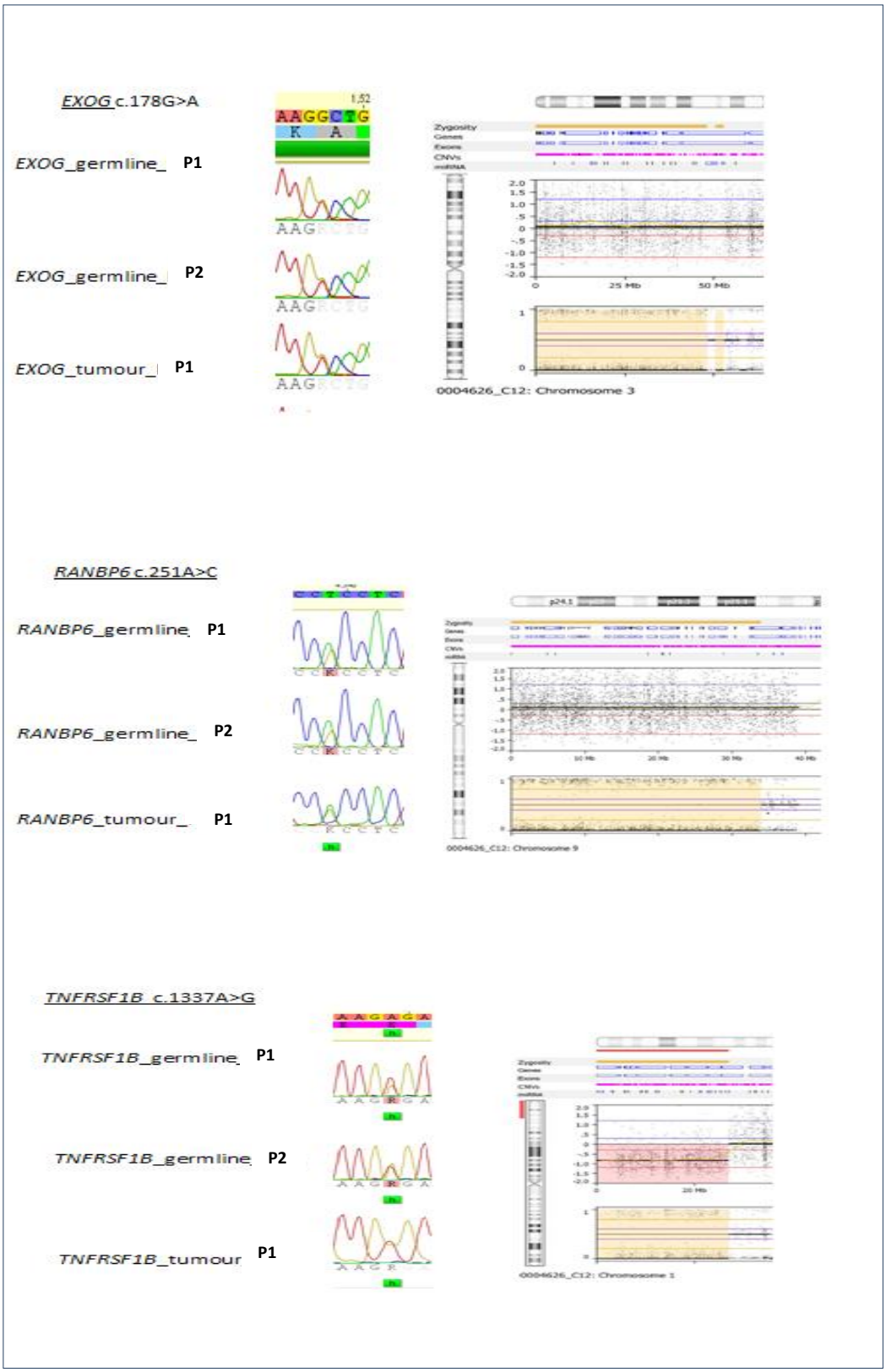
**Supplementary Data 6. Number of variants at each step of the filtering process**

| Sample ID | Non synonymous variants with an allele frequency 0.15 - 0.80 | Variants not seen in 1000G or NHLBI GO ESP | Variants post local exomes filter(n=147) | Variants shared by both cases | Shared variants excluding missense variants with a CADD score<10 | Sanger sequenced validated variants |
|---|---|---|---|---|---|---|
| **P1** | 4893 | 586 | 106 | 40 | 24 | 17 |
| **P2** | 4973 | 678 | 110 | | | |

1000G,1000 Genomes Project. NHLBI GO ESP, National Heart, Lung and Blood Institute Exome Sequencing Project.




**Supplementary Data 7. Sanger sequencing traces for the *RANBP2* missense variant in the germline of P1 (top left), germline of P2 (middle left), and tumour of P1 (bottom left), showing loss of the mutant allele.**




**Supplementary Data 8. Sanger sequencing traces for missense variants in *EXOG* (top panel), *RANBP6* (middle panel) and *TNFRSF1B* (bottom panel) showing the presence of the variant in the germline of P1, P2 and in tumour from P1, despite the presence of loss of heterozygosity of approximately a)50Mbs b)30Mbs and c)loss of heterozygosity and copy number loss of approximately 20Mbs.**

## Appendix 2. Sequencing performance summaries for validation cohort

### Table A-1. Sequencing performance summaries for the validation cohort of 337 samples

| Sample | Total reads | % Reads mapped | % Reads OnTarget | % Target bases >=10-fold Coverage | % Target bases >=20-fold Coverage | Mean coverage for target bases | Median Fragment Length |
|---|---|---|---|---|---|---|---|
| **CRC-WA-5371** | 7750722 | 99.72 | 87.01 | 97.88 | 97.52 | 904.95 | 138 |
| **CRC-WA-3953** | 5909066 | 99.35 | 87.87 | 98.02 | 97.51 | 691.57 | 130 |
| **CRC-MMC-5270** | 5190332 | 99.69 | 87.97 | 97.81 | 97.34 | 608.63 | 130 |
| **CRC-FCC-5507** | 4689494 | 99.30 | 87.66 | 97.81 | 97.32 | 541.18 | 126 |
| **CRC-KCF-01-005-0550** | 4539964 | 99.54 | 87.64 | 97.84 | 97.30 | 532.15 | 132 |
| **CRC-PMCC-4697** | 4546578 | 99.62 | 87.29 | 97.80 | 97.30 | 530.31 | 133 |
| **CRC-KCF-01-005-0555** | 4659658 | 99.41 | 87.46 | 97.83 | 97.28 | 541.73 | 130 |
| **CRC-NSW-22386** | 6009834 | 99.04 | 88.61 | 97.78 | 97.22 | 706.80 | 129 |
| **CRC-KCF-04-003-0099** | 4433576 | 99.41 | 88.39 | 97.77 | 97.19 | 521.13 | 127 |
| **CRC-KCF-02-005-0340** | 4122404 | 99.39 | 87.84 | 97.78 | 97.17 | 481.73 | 130 |
| **CRC-WA-5510** | 4952098 | 99.48 | 87.29 | 97.60 | 97.13 | 575.23 | 132 |
| **tablCRC-KCF-02-006-0347** | 4011658 | 99.30 | 87.48 | 97.79 | 97.12 | 465.86 | 129 |
| **CRC-KCF-01-008-0309** | 4095802 | 99.48 | 87.66 | 97.72 | 97.09 | 476.20 | 129 |
| **CRC-KCF-99-006-2243** | 4228558 | 99.56 | 88.09 | 97.65 | 97.09 | 495.81 | 128 |
| **CRC-KCF-03-007-0111** | 4228202 | 99.32 | 87.64 | 97.66 | 97.06 | 492.57 | 131 |
| **CRC-RMH-5542** | 3890668 | 99.40 | 87.59 | 97.78 | 97.06 | 453.30 | 130 |
| **CRC-KCF-99-004-1882** | 4276926 | 99.32 | 87.76 | 97.70 | 97.05 | 497.78 | 128 |
| **CRC-KCF-99-005-2499** | 4435112 | 99.49 | 87.73 | 97.72 | 97.05 | 519.74 | 131 |
| **CRC-MMC-5272** | 4263306 | 99.52 | 88.36 | 97.61 | 97.05 | 500.72 | 127 |
| **CRC-WA-5557** | 3950490 | 99.34 | 87.95 | 97.70 | 97.03 | 460.46 | 128 |

**Table A-1 (continued). Sequencing performance summaries for the validation cohort of 337 samples**

| Sample | Total reads | % Reads mapped | % Reads OnTarget | % Target bases >=10-fold Coverage | % Target bases >=20-fold Coverage | Mean coverage for target bases | Median Fragment Length |
|---|---|---|---|---|---|---|---|
| CRC-KCF-99-003-4146 | 4081388 | 99.63 | 87.40 | 97.68 | 97.02 | 476.39 | 133 |
| CRC-KCF-99-006-0409 | 4295738 | 99.30 | 87.92 | 97.66 | 97.02 | 503.78 | 131 |
| CRC-KCF-03-005-1534 | 4289564 | 99.59 | 87.44 | 97.63 | 97.01 | 500.11 | 131 |
| CRC-VCB-05PM0609 | 6967648 | 98.99 | 87.93 | 97.56 | 97.00 | 816.48 | 131 |
| CRC-FCC-6010 | 4017636 | 99.58 | 87.26 | 97.76 | 96.99 | 467.08 | 131 |
| CRC-KCF-03-005-1526 | 3674844 | 99.48 | 88.29 | 97.68 | 96.97 | 430.22 | 126 |
| CRC-KCF-12-005-235 | 4537040 | 99.34 | 87.96 | 97.60 | 96.97 | 531.04 | 129 |
| CRC-WA-5561 | 3695560 | 99.43 | 87.52 | 97.62 | 96.97 | 430.94 | 131 |
| CRC-KCF-07-007-0035 | 3529360 | 99.50 | 87.79 | 97.59 | 96.93 | 412.45 | 130 |
| CRC-FCC-5438 | 4815714 | 99.54 | 87.59 | 97.60 | 96.92 | 555.37 | 127 |
| CRC-MMC-5269 | 3901456 | 99.35 | 88.24 | 97.59 | 96.92 | 457.26 | 127 |
| CRC-KCF-00-004-0432 | 4503652 | 99.41 | 86.81 | 97.47 | 96.91 | 522.87 | 136 |
| CRC-NSW-16166 | 4401958 | 98.97 | 89.22 | 97.63 | 96.90 | 517.66 | 123 |
| CRC-4081-000 | 3460214 | 99.41 | 88.52 | 97.58 | 96.89 | 407.87 | 128 |
| CRC-KCF-05-001-2694 | 3556588 | 99.38 | 87.87 | 97.61 | 96.88 | 414.38 | 128 |
| CRC-KCF-00-003-0411 | 3986620 | 99.37 | 87.96 | 97.61 | 96.87 | 464.53 | 127 |
| CRC-KCF-04-006-1088 | 3689852 | 99.41 | 87.93 | 97.55 | 96.87 | 430.75 | 128 |
| CRC-KCF-13-002-91 | 3643534 | 99.46 | 88.03 | 97.59 | 96.86 | 425.34 | 127 |
| CRC-VCB-11PM1547 | 5872454 | 99.27 | 89.82 | 97.53 | 96.86 | 695.88 | 120 |
| CRC-KCF-05-001-3587 | 3586574 | 99.56 | 87.54 | 97.69 | 96.85 | 418.12 | 131 |
| CRC-KCF-99-004-1837 | 3449066 | 99.50 | 87.62 | 97.67 | 96.85 | 401.59 | 130 |
| CRC-FCC-5978 | 4804098 | 99.74 | 87.99 | 97.56 | 96.84 | 566.02 | 131 |

**Table A-1 (continued). Sequencing performance summaries for the validation cohort of 337 samples**

| Sample | Total reads | % Reads mapped | % Reads OnTarget | % Target bases >=10-fold Coverage | % Target bases >=20-fold Coverage | Mean coverage for target bases | Median Fragment Length |
|---|---|---|---|---|---|---|---|
| CRC-VCB-06RMH327 | 3893574 | 98.95 | 87.95 | 97.59 | 96.84 | 451.48 | 126 |
| CRC-FCC-5905 | 4307858 | 99.22 | 88.81 | 97.50 | 96.83 | 502.83 | 120 |
| CRC-KCF-00-002-0205 | 3771256 | 99.31 | 88.51 | 97.58 | 96.83 | 441.50 | 125 |
| CRC-4277-000 | 3916422 | 99.44 | 88.09 | 97.60 | 96.82 | 462.56 | 133 |
| CRC-KCF-03-007-0429 | 4181378 | 99.59 | 88.28 | 97.65 | 96.82 | 494.11 | 130 |
| CRC-VCB-08RMH0638 | 3724850 | 98.88 | 88.01 | 97.55 | 96.81 | 432.44 | 126 |
| CRC-RMH-5543 | 3467198 | 99.23 | 87.82 | 97.59 | 96.80 | 404.03 | 129 |
| CRC-VCB-11NH0228 | 4071348 | 98.92 | 88.56 | 97.62 | 96.80 | 476.51 | 122 |
| CRC-VCB-11AH0665 | 4030742 | 98.89 | 87.34 | 97.57 | 96.79 | 464.81 | 128 |
| CRC-WA-5851 | 3476178 | 99.26 | 88.10 | 97.59 | 96.78 | 403.60 | 125 |
| CRC-KCF-07-006-0213 | 3885260 | 99.40 | 88.58 | 97.53 | 96.77 | 456.17 | 124 |
| CRC-KCF-01-007-0359 | 3723030 | 99.34 | 88.57 | 97.53 | 96.75 | 436.24 | 124 |
| CRC-VCB-13AH0229 | 3461884 | 99.01 | 88.03 | 97.53 | 96.74 | 401.51 | 125 |
| CRC-MMC-5307 | 3339694 | 99.29 | 88.30 | 97.52 | 96.69 | 389.38 | 124 |
| CRC-FCC-5906 | 4042630 | 99.04 | 88.33 | 97.47 | 96.68 | 468.33 | 123 |
| CRC-4158-000 | 5027376 | 99.55 | 89.26 | 97.36 | 96.62 | 599.60 | 126 |
| CRC-VCB-11WH0060 | 3336146 | 99.04 | 88.48 | 97.44 | 96.62 | 390.21 | 126 |
| CRC-VCB-10NH230 | 3380794 | 99.01 | 88.30 | 97.53 | 96.60 | 394.55 | 125 |
| CRC-VCB-10WH0176 | 3256318 | 98.91 | 88.81 | 97.59 | 96.58 | 381.26 | 124 |
| CRC-WA-5548 | 3399200 | 99.23 | 87.91 | 97.49 | 96.58 | 395.63 | 128 |
| CRC-3950-000 | 3106908 | 99.45 | 88.55 | 97.47 | 96.57 | 366.88 | 129 |
| CRC-KCF-05-003-0424 | 2968510 | 99.21 | 87.40 | 97.44 | 96.57 | 342.32 | 127 |
| CRC-VCB-07AH578 | 3469804 | 98.88 | 88.87 | 97.48 | 96.55 | 404.18 | 121 |

**Table A-1 (continued). Sequencing performance summaries for the validation cohort of 337 samples**

| Sample | Total reads | % Reads mapped | % Reads OnTarget | % Target bases >=10-fold Coverage | % Target bases >=20-fold Coverage | Mean coverage for target bases | Median Fragment Length |
|---|---|---|---|---|---|---|---|
| CRC-NSW-26686 | 3397870 | 98.82 | 88.70 | 97.41 | 96.54 | 397.01 | 125 |
| CRC-VCB-04TB146 | 3029096 | 98.96 | 88.38 | 97.48 | 96.54 | 352.95 | 125 |
| CRC-3921-000 | 3157214 | 99.49 | 88.26 | 97.40 | 96.53 | 373.61 | 132 |
| CRC-VCB-05WH156 | 3296430 | 98.91 | 88.41 | 97.49 | 96.53 | 384.30 | 125 |
| CRC-VCB-10NH204 | 2924406 | 98.87 | 88.52 | 97.37 | 96.48 | 340.55 | 124 |
| CRC-VCB-12MH0494 | 3178904 | 98.98 | 89.08 | 97.42 | 96.48 | 374.79 | 123 |
| CRC-4278-000 | 3330764 | 99.41 | 88.46 | 97.36 | 96.47 | 391.83 | 127 |
| CRC-VCB-08WH0154 | 3446740 | 99.07 | 88.91 | 97.44 | 96.47 | 404.00 | 122 |
| CRC-3838-000 | 7042632 | 99.54 | 89.55 | 97.23 | 96.46 | 838.32 | 126 |
| CRC-4079-000 | 3231598 | 99.39 | 88.77 | 97.35 | 96.46 | 382.57 | 127 |
| CRC-KCF-04-005-0715 | 3319640 | 99.39 | 88.77 | 97.30 | 96.46 | 390.64 | 124 |
| CRC-3949-000 | 3110256 | 99.44 | 88.46 | 97.38 | 96.45 | 366.15 | 129 |
| CRC-NSW-22161 | 3359864 | 98.90 | 89.25 | 97.46 | 96.45 | 395.30 | 123 |
| CRC-VCB-04BL067 | 3057830 | 98.83 | 87.87 | 97.37 | 96.45 | 352.53 | 125 |
| CRC-VCB-05PM2162 | 3322620 | 98.96 | 88.90 | 97.40 | 96.45 | 389.54 | 122 |
| CRC-VCB-11SH100 | 4218574 | 99.02 | 89.51 | 97.46 | 96.45 | 504.87 | 123 |
| CRC-RMH-5335 | 3879482 | 99.29 | 89.94 | 97.34 | 96.42 | 456.47 | 114 |
| CRC-4078-000 | 3293952 | 99.41 | 88.41 | 97.37 | 96.41 | 388.04 | 129 |
| CRC-4162-000 | 2576508 | 99.47 | 88.35 | 97.39 | 96.41 | 303.66 | 129 |
| CRC-4280-000 | 3715964 | 99.52 | 89.06 | 97.38 | 96.41 | 442.85 | 127 |
| CRC-NSW-9213 | 4058972 | 99.03 | 90.42 | 97.44 | 96.41 | 480.15 | 114 |
| CRC-VCB-07AH0002 | 4967986 | 99.53 | 89.96 | 97.35 | 96.41 | 591.54 | 119 |
| CRC-VCB-06RMH209 | 3114914 | 99.07 | 88.49 | 97.45 | 96.40 | 363.24 | 122 |
| CRC-3954-000 | 2800094 | 99.45 | 88.73 | 97.26 | 96.39 | 330.56 | 127 |
| CRC-4426-000 | 2872882 | 99.44 | 88.78 | 97.29 | 96.39 | 340.89 | 129 |

**Table A-1 (continued). Sequencing performance summaries for the validation cohort of 337 samples**

| Sample | Total reads | % Reads mapped | % Reads OnTarget | % Target bases >=10-fold Coverage | % Target bases >=20-fold Coverage | Mean coverage for target bases | Median Fragment Length |
|---|---|---|---|---|---|---|---|
| CRC-NSW-100966 | 3572218 | 99.06 | 89.55 | 97.35 | 96.35 | 423.50 | 123 |
| CRC-3937-000 | 3026664 | 99.47 | 88.32 | 97.35 | 96.34 | 357.43 | 130 |
| CRC-4083-000 | 3180138 | 99.47 | 88.88 | 97.31 | 96.34 | 376.64 | 125 |
| CRC-4553-000 | 3292596 | 99.44 | 88.53 | 97.29 | 96.34 | 388.13 | 128 |
| CRC-NSW-16855 | 3214824 | 99.05 | 90.19 | 97.36 | 96.34 | 378.91 | 117 |
| CRC-4084-000 | 3832666 | 99.54 | 88.52 | 97.28 | 96.31 | 453.24 | 130 |
| CRC-VCB-11SH340 | 3714576 | 98.93 | 88.82 | 97.43 | 96.31 | 436.86 | 122 |
| CRC-NSW-22159 | 3329690 | 99.10 | 89.87 | 97.38 | 96.30 | 393.16 | 118 |
| CRC-VCB-10PM0148 | 3283550 | 99.43 | 86.09 | 97.34 | 96.29 | 377.30 | 134 |
| CRC-VCB-08WH172 | 2896176 | 98.90 | 88.39 | 97.34 | 96.28 | 336.69 | 124 |
| CRC-VCB-11SH086 | 3285472 | 98.82 | 87.52 | 97.37 | 96.28 | 378.72 | 125 |
| CRC-4655-000 | 2740434 | 99.47 | 87.81 | 97.31 | 96.27 | 320.75 | 132 |
| CRC-NSW-22382 | 2851232 | 98.88 | 89.14 | 97.36 | 96.27 | 333.77 | 122 |
| CRC-4197-000 | 3866236 | 99.48 | 88.75 | 97.24 | 96.25 | 457.63 | 126 |
| CRC-VCB-05TB067 | 2875030 | 98.92 | 88.70 | 97.39 | 96.25 | 334.09 | 121 |
| CRC-VCB-06PM1417 | 2554686 | 99.40 | 88.16 | 97.34 | 96.25 | 301.20 | 131 |
| CRC-4182-000 | 2861990 | 99.38 | 88.80 | 97.31 | 96.24 | 337.08 | 125 |
| CRC-NSW-100057 | 3151228 | 98.89 | 89.35 | 97.31 | 96.24 | 369.35 | 120 |
| CRC-NSW-4490 | 3289672 | 98.95 | 89.91 | 97.32 | 96.23 | 387.77 | 119 |
| CRC-3951-000 | 2973026 | 99.40 | 88.79 | 97.29 | 96.22 | 350.56 | 125 |
| CRC-4159-000 | 2590522 | 99.48 | 88.50 | 97.26 | 96.21 | 306.01 | 128 |
| CRC-VCB-00PM0230 | 2937628 | 98.88 | 89.17 | 97.24 | 96.21 | 342.74 | 118 |
| CRC-VCB-10SH125 | 3498822 | 98.81 | 88.50 | 97.42 | 96.20 | 405.85 | 122 |
| CRC-3257 | 3415864 | 99.73 | 90.07 | 97.21 | 96.19 | 407.02 | 118 |
| CRC-NSW-11603 | 3362644 | 99.01 | 89.76 | 97.32 | 96.19 | 396.87 | 119 |

**Table A-1 (continued). Sequencing performance summaries for the validation cohort of 337 samples**

| Sample | Total reads | % Reads mapped | % Reads OnTarget | % Target bases >=10-fold Coverage | % Target bases >=20-fold Coverage | Mean coverage for target bases | Median Fragment Length |
|---|---|---|---|---|---|---|---|
| CRC-VCB-10-00130 | 4068332 | 99.59 | 89.79 | 97.18 | 96.16 | 484.45 | 121 |
| CRC-4279-000 | 2778588 | 99.39 | 89.02 | 97.18 | 96.15 | 329.70 | 126 |
| CRC-VCB-05PM0412 | 2874716 | 99.05 | 89.23 | 97.37 | 96.15 | 337.35 | 121 |
| CRC-VCB-13WH0031 | 3054154 | 98.91 | 89.20 | 97.27 | 96.15 | 357.64 | 120 |
| CRC-4281-000 | 2746700 | 99.47 | 88.62 | 97.19 | 96.14 | 323.75 | 127 |
| CRC-4407-000 | 2470774 | 99.36 | 88.64 | 97.21 | 96.14 | 291.39 | 127 |
| CRC-VCB-11SH425 | 3504598 | 99.70 | 89.51 | 97.17 | 96.13 | 416.06 | 122 |
| CRC-4676-000 | 2954230 | 99.45 | 88.35 | 97.07 | 96.11 | 347.54 | 129 |
| CRC-NSW-23125 | 2923330 | 98.84 | 89.39 | 97.23 | 96.08 | 342.12 | 120 |
| CRC-VCB-06RMH224 | 2942808 | 98.70 | 88.53 | 97.25 | 96.08 | 340.97 | 122 |
| CRC-VCB-05NH109 | 3039814 | 99.65 | 89.44 | 97.17 | 96.07 | 360.97 | 122 |
| CRC-FCC-4284 | 2745448 | 99.63 | 88.34 | 97.24 | 96.06 | 319.50 | 122 |
| CRC-NSW-11629 | 3197026 | 98.89 | 90.37 | 97.21 | 96.06 | 377.05 | 115 |
| CRC-NSW-134486 | 2924514 | 99.61 | 89.51 | 97.11 | 96.06 | 346.82 | 122 |
| CRC-4318-000 | 2578834 | 99.48 | 88.86 | 97.31 | 96.04 | 304.85 | 126 |
| CRC-FCC-5545 | 2820418 | 99.48 | 87.82 | 97.16 | 96.04 | 325.24 | 125 |
| CRC-NSW-9094 | 2736994 | 98.80 | 89.13 | 97.20 | 96.04 | 319.68 | 121 |
| CRC-4085-000 | 3166842 | 99.48 | 88.70 | 97.06 | 96.03 | 373.87 | 127 |
| CRC-VCB-03PM0333 | 2954296 | 98.90 | 88.40 | 97.34 | 96.02 | 343.49 | 122 |
| CRC-3839-000 | 2481352 | 99.37 | 88.32 | 97.25 | 95.97 | 291.69 | 129 |
| CRC-VCB-15PM1014 | 2345718 | 99.44 | 87.14 | 97.17 | 95.97 | 273.14 | 133 |
| CRC-VCB-13SH143 | 2880250 | 99.02 | 88.63 | 97.21 | 95.95 | 335.13 | 122 |
| CRC-3940-000 | 2440436 | 99.43 | 88.56 | 97.10 | 95.94 | 287.06 | 127 |
| CRC-VCB-2008-01313 | 3644418 | 99.54 | 89.92 | 96.89 | 95.94 | 433.73 | 120 |
| CRC-4583-000 | 2512134 | 99.41 | 87.87 | 97.14 | 95.91 | 293.63 | 132 |

**Table A-1 (continued). Sequencing performance summaries for the validation cohort of 337 samples**

| Sample | Total reads | % Reads mapped | % Reads OnTarget | % Target bases >=10-fold Coverage | % Target bases >=20-fold Coverage | Mean coverage for target bases | Median Fragment Length |
|---|---|---|---|---|---|---|---|
| CRC-4080-000 | 2542370 | 99.48 | 88.74 | 97.02 | 95.90 | 300.63 | 127 |
| CRC-4467-000 | 2505086 | 99.52 | 88.62 | 97.05 | 95.90 | 295.76 | 127 |
| CRC-4362-000 | 2345070 | 99.40 | 88.57 | 97.23 | 95.88 | 277.08 | 129 |
| CRC-VCB-11PM0135 | 3338926 | 99.63 | 90.06 | 96.98 | 95.88 | 397.13 | 118 |
| CRC-VCB-11SH726 | 2780822 | 98.73 | 89.27 | 97.15 | 95.88 | 325.63 | 121 |
| CRC-VCB-11PM0197 | 4102250 | 99.77 | 90.45 | 96.81 | 95.86 | 493.65 | 120 |
| CRC-VCB-09WH0117 | 3022004 | 99.71 | 89.65 | 96.92 | 95.85 | 359.06 | 120 |
| CRC-3881-000 | 2198376 | 99.42 | 88.11 | 97.07 | 95.83 | 258.68 | 131 |
| CRC-VCB-09CH050 | 3278970 | 99.67 | 89.71 | 97.00 | 95.83 | 390.14 | 121 |
| CRC-VCB-09PM0930 | 2266782 | 99.36 | 87.35 | 97.08 | 95.82 | 266.40 | 137 |
| CRC-4086-000 | 2442416 | 99.34 | 88.89 | 96.95 | 95.81 | 288.14 | 126 |
| CRC-VCB-10-01342 | 3246148 | 99.73 | 89.90 | 96.85 | 95.81 | 387.26 | 120 |
| CRC-VCB-11SH672 | 3265372 | 99.00 | 87.09 | 97.12 | 95.81 | 376.18 | 128 |
| CRC-VCB-2009-00532 | 3003316 | 99.67 | 89.61 | 96.94 | 95.81 | 356.58 | 121 |
| CRC-3805-000 | 2600464 | 99.47 | 88.81 | 97.06 | 95.79 | 306.85 | 125 |
| CRC-NSW-45803 | 2395148 | 98.80 | 89.14 | 97.14 | 95.78 | 279.66 | 121 |
| CRC-3938-002 | 2408320 | 99.31 | 89.06 | 96.99 | 95.76 | 284.16 | 122 |
| CRC-NSW-11332 | 3397212 | 99.04 | 91.30 | 97.17 | 95.76 | 402.67 | 108 |
| CRC-VCB-09PM0932 | 3825062 | 99.67 | 90.45 | 96.75 | 95.76 | 458.30 | 117 |
| CRC-VCB-10-00793 | 2471784 | 99.64 | 89.71 | 96.96 | 95.75 | 293.78 | 121 |
| CRC-VCB-10PM0143 | 2283570 | 99.39 | 86.54 | 97.10 | 95.75 | 263.19 | 130 |
| CRC-VCB-05WH147 | 3085434 | 99.71 | 89.94 | 97.07 | 95.74 | 367.14 | 120 |
| CRC-VCB-10-01030 | 3040300 | 99.59 | 89.81 | 96.82 | 95.74 | 361.07 | 120 |
| CRC-1806 | 3018410 | 99.61 | 90.05 | 96.87 | 95.73 | 358.78 | 119 |

**Table A-1 (continued). Sequencing performance summaries for the validation cohort of 337 samples**

| Sample | Total reads | % Reads mapped | % Reads OnTarget | % Target bases >=10-fold Coverage | % Target bases >=20-fold Coverage | Mean coverage for target bases | Median Fragment Length |
|---|---|---|---|---|---|---|---|
| CRC-VCB-10PM2232 | 2811104 | 99.36 | 88.52 | 96.94 | 95.71 | 331.12 | 127 |
| CRC-VCB-11PM0292 | 2649746 | 99.63 | 89.70 | 96.92 | 95.71 | 314.84 | 120 |
| CRC-NSW-125424 | 2566934 | 99.52 | 89.97 | 96.83 | 95.70 | 304.38 | 117 |
| CRC-4076-000 | 2206954 | 99.44 | 88.11 | 97.02 | 95.69 | 260.25 | 132 |
| CRC-VCB-07PM0091 | 2471102 | 99.39 | 88.50 | 97.12 | 95.69 | 293.89 | 130 |
| CRC-VCB-2007-00984 | 3325452 | 99.62 | 90.03 | 96.80 | 95.69 | 395.77 | 119 |
| CRC-VCB-09SH801 | 3398820 | 99.61 | 89.89 | 96.90 | 95.68 | 404.95 | 120 |
| CRC-4328-000 | 2034332 | 99.42 | 88.39 | 96.94 | 95.67 | 240.74 | 131 |
| CRC-VCB-10-01072 | 3133270 | 99.57 | 89.90 | 96.78 | 95.66 | 372.46 | 119 |
| CRC-VCB-05PM2155 | 2369970 | 99.73 | 89.92 | 96.85 | 95.65 | 282.57 | 120 |
| CRC-VCB-12PM0086 | 2503800 | 99.73 | 89.86 | 96.87 | 95.65 | 298.53 | 120 |
| CRC-VCB-2007-00780 | 2704850 | 99.68 | 89.93 | 96.92 | 95.65 | 322.76 | 120 |
| CRC-VCB-2008-01986 | 2921712 | 99.62 | 90.02 | 96.82 | 95.65 | 347.03 | 117 |
| CRC-NSW-125733 | 3079412 | 99.71 | 89.97 | 96.89 | 95.64 | 368.44 | 120 |
| CRC-NSW-22333 | 2621978 | 98.84 | 89.56 | 96.99 | 95.64 | 308.63 | 121 |
| CRC-VCB-08PM2111 | 2640770 | 99.39 | 89.16 | 96.98 | 95.64 | 314.95 | 126 |
| CRC-NSW-10848 | 2719474 | 99.04 | 90.50 | 96.96 | 95.63 | 321.96 | 114 |
| CRC-VCB-06AH391 | 2237132 | 98.92 | 88.46 | 97.10 | 95.62 | 260.24 | 123 |
| CRC-NSW-21602 | 2940434 | 98.93 | 90.58 | 96.97 | 95.61 | 348.70 | 114 |
| CRC-NSW-11439 | 2468422 | 98.85 | 89.80 | 96.90 | 95.60 | 289.86 | 117 |
| CRC-3952-000 | 2649866 | 99.44 | 89.16 | 96.92 | 95.59 | 314.19 | 124 |
| CRC-4270-000 | 3285434 | 99.48 | 89.80 | 97.06 | 95.59 | 391.21 | 121 |
| CRC-VCB-07PM1143 | 2357024 | 99.41 | 88.74 | 96.91 | 95.59 | 278.82 | 126 |

**Table A-1 (continued). Sequencing performance summaries for the validation cohort of 337 samples**

| Sample | Total reads | % Reads mapped | % Reads OnTarget | % Target bases >=10-fold Coverage | % Target bases >=20-fold Coverage | Mean coverage for target bases | Median Fragment Length |
|---|---|---|---|---|---|---|---|
| CRC-4554-000 | 2441964 | 99.43 | 88.90 | 96.91 | 95.57 | 288.28 | 126 |
| CRC-VCB-2005-02299 | 2589252 | 99.76 | 89.80 | 96.90 | 95.57 | 308.63 | 121 |
| CRC-VCB-10AH774 | 2445942 | 99.08 | 88.19 | 97.09 | 95.56 | 284.08 | 123 |
| CRC-VCB-06PM0112 | 2558300 | 99.60 | 89.44 | 96.95 | 95.55 | 303.19 | 122 |
| CRC-VCB-12PM0217 | 3080404 | 99.60 | 90.08 | 96.73 | 95.55 | 367.17 | 118 |
| CRC-4163-000 | 2496002 | 99.29 | 88.94 | 96.78 | 95.53 | 294.38 | 124 |
| CRC-NSW-11984 | 2945926 | 98.98 | 91.49 | 96.97 | 95.52 | 350.23 | 109 |
| CRC-VCB-11PM0829 | 2514382 | 99.41 | 88.35 | 97.08 | 95.51 | 301.50 | 135 |
| CRC-NSW-134541 | 2582862 | 99.64 | 89.61 | 96.67 | 95.50 | 307.03 | 121 |
| CRC-4082-000 | 3111060 | 99.45 | 89.36 | 96.87 | 95.49 | 370.47 | 123 |
| CRC-4545-000 | 2472162 | 99.50 | 89.75 | 96.76 | 95.49 | 293.31 | 120 |
| CRC-VCB-07RMH094 | 2436308 | 99.62 | 89.82 | 96.72 | 95.49 | 290.00 | 121 |
| CRC-4282-000 | 2040318 | 99.30 | 88.70 | 96.99 | 95.48 | 239.80 | 126 |
| CRC-NSW-134275 | 3212816 | 99.73 | 90.32 | 96.82 | 95.48 | 385.15 | 119 |
| CRC-3235 | 2818992 | 99.64 | 89.86 | 96.79 | 95.47 | 335.38 | 120 |
| CRC-4160-000 | 2059896 | 99.26 | 88.45 | 96.91 | 95.47 | 241.08 | 126 |
| CRC-NSW-7412 | 3154520 | 99.74 | 90.06 | 96.79 | 95.47 | 375.31 | 119 |
| CRC-2958-000 | 2724540 | 99.46 | 90.03 | 96.77 | 95.46 | 323.72 | 117 |
| CRC-3842-000 | 3583394 | 99.50 | 90.20 | 96.65 | 95.46 | 432.89 | 124 |
| CRC-4386-000 | 2559144 | 99.41 | 89.42 | 96.86 | 95.46 | 301.81 | 121 |
| CRC-3448 | 2442464 | 99.71 | 89.53 | 96.83 | 95.45 | 290.25 | 122 |
| CRC-NSW-131016 | 7887026 | 99.78 | 92.56 | 96.49 | 95.45 | 961.77 | 109 |
| CRC-NSW-134267 | 3165666 | 99.58 | 90.07 | 96.51 | 95.44 | 377.51 | 119 |
| CRC-VCB-16PM0214 | 2794060 | 99.63 | 89.89 | 96.81 | 95.44 | 331.94 | 119 |
| CRC-NSW-11929 | 2845434 | 98.88 | 91.41 | 96.89 | 95.41 | 337.44 | 109 |

**Table A-1 (continued). Sequencing performance summaries for the validation cohort of 337 samples**

| Sample | Total reads | % Reads mapped | % Reads OnTarget | % Target bases >=10-fold Coverage | % Target bases >=20-fold Coverage | Mean coverage for target bases | Median Fragment Length |
|---|---|---|---|---|---|---|---|
| CRC-NSW-134123 | 2979578 | 99.74 | 90.06 | 96.63 | 95.40 | 355.98 | 121 |
| CRC-VCB-10SH566 | 2503938 | 99.72 | 89.82 | 96.65 | 95.39 | 298.13 | 120 |
| CRC-NSW-11463 | 3414184 | 98.90 | 91.86 | 96.82 | 95.38 | 407.17 | 107 |
| CRC-NSW-11555 | 2975218 | 98.90 | 91.68 | 96.76 | 95.38 | 353.25 | 108 |
| CRC-NSW-22343 | 2781394 | 98.66 | 90.22 | 96.74 | 95.38 | 328.19 | 117 |
| CRC-VCB-06AH036 | 2160390 | 98.82 | 88.48 | 96.97 | 95.35 | 250.01 | 122 |
| CRC-4236-000 | 2063044 | 99.32 | 88.63 | 96.71 | 95.34 | 242.25 | 125 |
| CRC-VCB-06PM1298 | 2165688 | 99.48 | 88.92 | 97.00 | 95.34 | 257.57 | 126 |
| CRC-VCB-11PM0721 | 2555062 | 99.69 | 89.71 | 96.72 | 95.34 | 303.79 | 120 |
| CRC-4276-000 | 2387078 | 99.44 | 89.08 | 96.89 | 95.32 | 282.32 | 125 |
| CRC-VCB-11PM0822 | 2581664 | 99.67 | 89.91 | 96.57 | 95.31 | 307.39 | 119 |
| CRC-VCB-11SH774 | 2426506 | 98.90 | 90.15 | 96.72 | 95.31 | 288.36 | 118 |
| CRC-NSW-13164 | 2657480 | 98.82 | 90.66 | 96.71 | 95.30 | 314.32 | 115 |
| CRC-4196-000 | 2762402 | 99.44 | 89.54 | 96.57 | 95.29 | 327.55 | 122 |
| CRC-VCB-10SH800 | 2572284 | 99.52 | 90.01 | 96.62 | 95.28 | 306.50 | 119 |
| CRC-4083-001 | 2322526 | 99.46 | 89.40 | 96.73 | 95.27 | 275.08 | 121 |
| CRC-NSW-15338 | 3465302 | 99.70 | 91.18 | 96.62 | 95.27 | 417.72 | 112 |
| CRC-VCB-11AH0342 | 2530350 | 99.54 | 89.64 | 96.71 | 95.26 | 299.07 | 118 |
| CRC-VCB-13WH0047 | 2449394 | 99.78 | 89.95 | 96.72 | 95.26 | 292.23 | 119 |
| CRC-NSW-11498 | 2543660 | 98.89 | 90.63 | 96.76 | 95.24 | 301.50 | 114 |
| CRC-VCB-10PM02139 | 2540882 | 99.53 | 89.97 | 96.63 | 95.24 | 301.71 | 118 |
| CRC-4161-000 | 2464312 | 99.40 | 88.82 | 96.77 | 95.23 | 292.65 | 128 |
| CRC-4269-000 | 2358194 | 99.46 | 89.56 | 96.68 | 95.23 | 279.28 | 121 |
| CRC-VCB-11PM0551 | 2208240 | 99.52 | 89.70 | 96.71 | 95.23 | 261.49 | 120 |

**Table A-1 (continued). Sequencing performance summaries for the validation cohort of 337 samples**

| Sample | Total reads | % Reads mapped | % Reads OnTarget | % Target bases >=10-fold Coverage | % Target bases >=20-fold Coverage | Mean coverage for target bases | Median Fragment Length |
|---|---|---|---|---|---|---|---|
| CRC-VCB-13SH040 | 2484780 | 99.50 | 89.76 | 96.64 | 95.21 | 294.68 | 120 |
| CRC-VCB-1999-00113 | 2819938 | 99.72 | 90.22 | 96.60 | 95.21 | 336.24 | 117 |
| CRC-VCB-13EH0118 | 2342110 | 99.03 | 88.40 | 96.88 | 95.20 | 273.22 | 124 |
| CRC-VCB-2007-00902 | 2411470 | 99.71 | 90.14 | 96.58 | 95.20 | 287.65 | 118 |
| CRC-VCB-2008-00236 | 2760040 | 99.68 | 89.95 | 96.61 | 95.19 | 328.52 | 119 |
| CRC-4226-000 | 2183192 | 99.28 | 89.41 | 96.66 | 95.18 | 257.27 | 121 |
| CRC-NSW-125421 | 2667136 | 99.72 | 90.26 | 96.61 | 95.18 | 318.81 | 117 |
| CRC-VCB-09PM0241 | 2027602 | 98.49 | 88.77 | 96.69 | 95.18 | 234.93 | 120 |
| CRC-NSW-11743 | 2376806 | 98.74 | 90.56 | 96.85 | 95.17 | 279.33 | 113 |
| CRC-NSW-20834 | 3385454 | 99.79 | 91.12 | 96.31 | 95.17 | 408.54 | 115 |
| CRC-VCB-10AH560 | 2190918 | 98.89 | 88.82 | 96.81 | 95.17 | 254.87 | 121 |
| CRC-VCB-11PM0593 | 2572886 | 99.67 | 89.92 | 96.70 | 95.15 | 306.20 | 118 |
| CRC-VCB-15PM0872 | 2564916 | 99.73 | 90.17 | 96.54 | 95.14 | 306.36 | 118 |
| CRC-VCB-12BH0020 | 2731454 | 98.78 | 90.63 | 96.74 | 95.13 | 326.26 | 114 |
| CRC-WA-4904 | 2099914 | 99.60 | 90.06 | 96.59 | 95.13 | 249.17 | 119 |
| CRC-VCB-10-00272 | 2199246 | 99.70 | 89.68 | 96.58 | 95.11 | 261.39 | 121 |
| CRC-VCB-07PM0660 | 2529722 | 99.34 | 87.68 | 96.71 | 95.10 | 297.83 | 132 |
| CRC-VCB-10PM1612 | 2217974 | 99.37 | 88.50 | 96.92 | 95.10 | 264.97 | 131 |
| CRC-VCB-11PM1483 | 2576944 | 99.72 | 90.11 | 96.64 | 95.08 | 308.16 | 119 |
| CRC-VCB-13AH0114 | 2331924 | 99.71 | 89.60 | 96.47 | 95.08 | 276.33 | 120 |
| CRC-VCB-13AL001 | 2146866 | 99.77 | 89.96 | 96.54 | 95.07 | 256.10 | 119 |
| CRC-VCB-11PM1300 | 2545404 | 99.43 | 89.91 | 96.50 | 95.06 | 301.75 | 119 |

**Table A-1 (continued). Sequencing performance summaries for the validation cohort of 337 samples**

| Sample | Total reads | % Reads mapped | % Reads OnTarget | % Target bases >=10-fold Coverage | % Target bases >=20-fold Coverage | Mean coverage for target bases | Median Fragment Length |
|---|---|---|---|---|---|---|---|
| CRC-VCB-13AL117 | 2415682 | 98.96 | 89.48 | 96.75 | 95.02 | 283.49 | 116 |
| CRC-NSW-7241 | 2334034 | 98.88 | 90.57 | 96.63 | 95.01 | 275.51 | 113 |
| CRC-VCB-10RMH0097 | 2582766 | 98.90 | 89.87 | 96.60 | 94.98 | 305.31 | 119 |
| CRC-VCB-10-00921 | 2016254 | 99.73 | 89.85 | 96.45 | 94.94 | 240.01 | 120 |
| CRC-VCB-15PM1010 | 2432908 | 99.66 | 89.91 | 96.27 | 94.92 | 289.83 | 120 |
| CRC-VCB-06PM1323 | 2287254 | 99.46 | 88.67 | 96.55 | 94.88 | 270.70 | 126 |
| CRC-VCB-15PM0121 | 1861964 | 99.40 | 88.29 | 96.65 | 94.88 | 218.46 | 126 |
| CRC-VCB-2007-01016 | 2429990 | 99.61 | 90.08 | 96.26 | 94.87 | 289.21 | 118 |
| CRC-4274-000 | 2410842 | 99.38 | 89.32 | 96.74 | 94.86 | 285.11 | 124 |
| CRC-NSW-15158 | 3019980 | 99.59 | 91.18 | 96.53 | 94.82 | 365.32 | 115 |
| CRC-VCB-10NH210 | 2005060 | 99.53 | 89.58 | 96.49 | 94.81 | 237.14 | 120 |
| CRC-VCB-2004-01379 | 2717830 | 99.73 | 90.63 | 96.21 | 94.78 | 324.14 | 113 |
| CRC-NSW-27347 | 2627952 | 98.96 | 91.01 | 96.56 | 94.77 | 313.58 | 113 |
| CRC-NSW-125897 | 2078338 | 99.65 | 89.79 | 96.39 | 94.75 | 246.04 | 118 |
| CRC-VCB-08RMH0267 | 2275180 | 98.88 | 90.39 | 96.71 | 94.75 | 267.46 | 111 |
| CRC-NSW-134450 | 2005428 | 99.66 | 89.88 | 96.32 | 94.73 | 238.79 | 120 |
| CRC-VCB-12NH0031 | 2346796 | 98.74 | 90.49 | 96.58 | 94.73 | 281.81 | 117 |
| CRC-4356 | 2077584 | 99.59 | 89.76 | 96.32 | 94.72 | 246.11 | 119 |
| CRC-VCB-10CH033 | 2342956 | 98.95 | 88.85 | 96.56 | 94.71 | 273.36 | 118 |
| CRC-VCB-10SH156 | 2056888 | 99.68 | 89.87 | 96.26 | 94.71 | 244.77 | 120 |
| CRC-VCB-10CH094 | 2760642 | 98.91 | 87.50 | 96.54 | 94.69 | 316.67 | 122 |
| CRC-NSW-125610 | 1986310 | 99.70 | 90.03 | 96.52 | 94.67 | 236.29 | 117 |
| CRC-NSW-15072 | 2127854 | 99.68 | 90.52 | 96.36 | 94.67 | 253.85 | 114 |
| CRC-KCF-02-006-1388 | 3444818 | 99.17 | 90.54 | 96.46 | 94.63 | 412.93 | 117 |

**Table A-1 (continued). Sequencing performance summaries for the validation cohort of 337 samples**

| Sample | Total reads | % Reads mapped | % Reads OnTarget | % Target bases >=10-fold Coverage | % Target bases >=20-fold Coverage | Mean coverage for target bases | Median Fragment Length |
|---|---|---|---|---|---|---|---|
| CRC-VCB-16PM0029 | 2134708 | 99.53 | 90.25 | 96.27 | 94.57 | 253.99 | 117 |
| CRC-3922-000 | 3301766 | 99.47 | 88.35 | 96.15 | 94.55 | 388.89 | 135 |
| CRC-VCB-12PM0812 | 2246808 | 99.54 | 90.17 | 96.13 | 94.47 | 267.31 | 118 |
| CRC-NSW-8585 | 2370666 | 98.52 | 91.40 | 96.23 | 94.45 | 278.03 | 107 |
| CRC-NSW-11434 | 1838162 | 98.48 | 90.58 | 96.23 | 94.25 | 214.55 | 111 |
| CRC-NSW-11365 | 2126484 | 98.53 | 91.20 | 96.09 | 94.23 | 249.66 | 108 |
| CRC-VCB-11EH0096 | 2519930 | 98.95 | 91.28 | 96.03 | 94.05 | 305.08 | 114 |
| CRC-4363-000 | 2252826 | 99.51 | 89.14 | 96.10 | 94.04 | 265.30 | 124 |
| CRC-VCB-2008-00260 | 1764146 | 99.57 | 90.00 | 96.00 | 94.01 | 208.71 | 117 |
| CRC-VCB-11SH459 | 2077194 | 99.51 | 90.16 | 95.95 | 94.00 | 245.20 | 115 |
| CRC-VCB-08PM2024 | 2109254 | 99.23 | 86.54 | 96.35 | 93.99 | 242.47 | 133 |
| CRC-VCB-13BH0040 | 1784186 | 99.74 | 90.72 | 95.78 | 93.96 | 213.72 | 114 |
| CRC-NSW-14252 | 2410652 | 99.76 | 91.75 | 95.78 | 93.95 | 293.14 | 113 |
| CRC-3948-000 | 1682828 | 99.17 | 89.16 | 96.03 | 93.90 | 198.17 | 122 |
| CRC-4406-000 | 2233860 | 99.37 | 87.61 | 96.13 | 93.90 | 258.81 | 133 |
| CRC-NSW-134447 | 1581466 | 99.75 | 89.90 | 95.95 | 93.90 | 188.04 | 118 |
| CRC-VCB-08PM1750 | 2163654 | 98.84 | 90.19 | 96.02 | 93.90 | 254.51 | 112 |
| CRC-VCB-08RMH0720 | 1592862 | 99.55 | 89.93 | 96.05 | 93.89 | 188.65 | 118 |
| CRC-3434-001 | 1852910 | 99.24 | 90.50 | 96.19 | 93.82 | 222.54 | 118 |
| CRC-NSW-13825 | 2289492 | 99.75 | 91.66 | 95.53 | 93.82 | 277.10 | 110 |
| CRC-NSW-14337 | 2014954 | 99.77 | 91.10 | 95.42 | 93.49 | 243.78 | 117 |
| CRC-4271-000 | 2417822 | 99.38 | 90.49 | 95.72 | 93.35 | 291.61 | 121 |
| CRC-VCB-10CH087 | 1607004 | 99.64 | 90.33 | 95.40 | 93.29 | 191.58 | 115 |
| CRC-VCB-10AH059 | 2190040 | 98.78 | 89.29 | 95.72 | 93.28 | 252.99 | 115 |

**Table A-1 (continued). Sequencing performance summaries for the validation cohort of 337 samples**

| Sample | Total reads | % Reads mapped | % Reads OnTarget | % Target bases >=10-fold Coverage | % Target bases >=20-fold Coverage | Mean coverage for target bases | Median Fragment Length |
|---|---|---|---|---|---|---|---|
| CRC-NSW-11446 | 2456396 | 98.54 | 92.44 | 95.35 | 92.72 | 295.14 | 106 |
| CRC-4272-000 | 2145116 | 99.45 | 90.33 | 95.36 | 92.63 | 258.07 | 121 |
| CRC-NSW-15003 | 2009870 | 99.75 | 91.54 | 95.25 | 92.56 | 243.74 | 112 |
| CRC-4275-000 | 2037010 | 99.26 | 90.37 | 95.44 | 92.39 | 243.61 | 118 |
| CRC-NSW-25902 | 1684352 | 99.61 | 91.17 | 95.01 | 92.32 | 203.17 | 114 |
| CRC-4283-000 | 2423304 | 99.27 | 93.91 | 94.43 | 91.21 | 291.05 | 94 |
| CRC-4273-000 | 2409470 | 99.30 | 91.48 | 94.36 | 90.71 | 293.31 | 117 |
| CRC-3544-000 | 2322570 | 99.32 | 91.05 | 93.60 | 89.78 | 281.59 | 119 |
| CRC-4285-000 | 4199028 | 99.56 | 89.35 | 91.90 | 86.69 | 495.66 | 129 |
| CRC-KCF-00-009-0341 | 4246388 | 99.20 | 89.02 | 97.65 | 96.84 | 501.28 | 122 |
| CRC-KCF-11-002-0218 | 3715744 | 99.36 | 88.19 | 97.65 | 96.84 | 435.52 | 128 |
| CRC-VCB-07RMH049 | 3189750 | 98.88 | 88.35 | 97.49 | 96.55 | 371.85 | 125 |
| CRC-4195-000 | 3929316 | 99.50 | 88.84 | 97.23 | 96.36 | 466.72 | 129 |
| CRC-3558-000 | 2725572 | 99.51 | 88.37 | 97.38 | 96.35 | 321.88 | 131 |
| CRC-3868-000 | 2858708 | 99.52 | 88.63 | 97.20 | 96.17 | 338.50 | 129 |
| CRC-VCB-05PM0797 | 3088796 | 98.89 | 88.83 | 97.36 | 96.17 | 361.55 | 123 |
| CRC-NSW-14628 | 2934904 | 98.97 | 89.62 | 97.09 | 95.91 | 345.90 | 120 |
| CRC-NSW-18945 | 2442678 | 98.94 | 89.15 | 97.26 | 95.91 | 285.98 | 122 |
| CRC-NSW-11419 | 3365372 | 99.03 | 90.95 | 97.07 | 95.72 | 400.98 | 113 |
| CRC-VCB-2009-00711 | 2873612 | 99.60 | 89.86 | 96.77 | 95.59 | 341.78 | 120 |
| CRC-NSW-13982 | 2923548 | 98.95 | 90.66 | 96.91 | 95.58 | 346.28 | 113 |
| CRC-VCB-07WH224 | 2396788 | 98.78 | 88.15 | 96.94 | 95.41 | 275.95 | 124 |
| CRC-4609-000 | 1919928 | 99.33 | 87.72 | 96.91 | 95.40 | 223.40 | 133 |
| CRC-VCB-11PM0248 | 2415528 | 99.40 | 86.50 | 96.85 | 95.21 | 277.88 | 128 |
| CRC-NSW-134229 | 2313188 | 99.59 | 89.96 | 96.52 | 95.05 | 274.91 | 118 |

**Table A-1 (continued). Sequencing performance summaries for the validation cohort of 337 samples**

| Sample | Total reads | % Reads mapped | % Reads OnTarget | % Target bases >=10-fold Coverage | % Target bases >=20-fold Coverage | Mean coverage for target bases | Median Fragment Length |
|---|---|---|---|---|---|---|---|
| CRC-VCB-11AH0082 | 2369242 | 99.72 | 89.98 | 96.52 | 95.03 | 282.84 | 119 |
| CRC-4590-000 | 3119144 | 99.43 | 87.34 | 96.49 | 94.62 | 361.52 | 137 |
| CRC-VCB-05TB023 | 1977750 | 99.68 | 89.54 | 96.37 | 94.61 | 235.04 | 121 |
| CRC-NSW-14461 | 3192536 | 98.74 | 92.78 | 95.69 | 93.21 | 389.69 | 109 |
| CRC-3882-000 | 2102382 | 99.40 | 89.99 | 95.67 | 93.10 | 252.77 | 124 |
| CRC-NSW-15001 | 2665778 | 99.75 | 91.49 | 95.11 | 92.86 | 323.03 | 114 |

**Appendix 3. Thirty genes with an ExAC pLI score>0.9 and a LoF variant in the discovery cohort**

**Table A-2 (continued). Thirty genes with an ExAC pLI score>0.9 and a LoF variant in the discovery cohort**

| Gene | pLI | Sample | Variant | Gene function from GeneCards or PubMed |
|---|---|---|---|---|
| *MYCBP2* | 1.00 | PC11 | ENST00000407578.2: c.4756C>T | Atypical E3 ubiquitin-protein ligase that specifically mediates ubiquitination of threonine and serine residues on target proteins |
| *EP300* | 1.00 | PC16 | ENST00000263253.7: c.907-2_907-1delAG | histone acetyltransferase that regulates transcription via chromatin remodelling |
| *HTT* | 1.00 | PC38 | ENST00000355072.5: c.118_119insAG | May play a role in microtubule-mediated transport or vesicle function |
| *ESPL1* | 1.00 | PC37 | ENST00000257934.4: c.2619+2T>C | caspase-like protease that initiates the final separation of sister chromatids during anaphase |
| *ARID4A* | 1.00 | PC48 | ENST00000355431.3: c.3211+1G>A | DNA-binding protein that modulates activity of several transcription factors including RB1 (retinoblastoma-associated protein) and AR (androgen receptor) (By similarity) |
| *PTPRA* | 1.00 | PC45 | ENST00000216877.6: c.1993C>T | dephosphorylates and activates Src tyrosine kinases |
| *RNF111* | 1.00 | PC1 | ENST00000557998.1: c.919_920insC | E3 ubiquitin ligase that ubiquinates inhibitory SMADs |
| *SHANK3* | 1.00 | PC11 | ENST00000262795.3: c.1331_1332insG | multidomain scaffold protein, linked to autism and schizophrenia |
| *KDM2B* | 1.00 | PC29 | ENST00000377071.4: c.83_84delAC | Histone demethylase that demethylates 'Lys-4' and 'Lys-36' of histone H3, thereby playing a central role in histone code |
| *MYEF2* | 1.00 | PC21 | ENST00000324324.7: c.1138+1G>T | Transcriptional repressor of the myelin basic protein gene |
| *SAMD4A* | 1.00 | PC9 | ENST00000555192.1: c.919C>T | Acts as a translational repressor of SRE-containing messengers |
| *KCNMA1* | 1.00 | PC28 | ENST00000404857.1: c.2433+1G>A | potassium channel |
| *CREBRF* | 1.00 | PC26 | ENST00000296953.2: c.-191-5_-191-2delGACA | negative regulator of endoplasmic reticulum stress response |
| *TNFSF12-TNFSF13* | 1.00 | PC27 | ENST00000293826.4: c.879_880insGCAGGTGAGAGCCGTGTGG | hybrid protein, stimulates cycling in B and T lymphoma cell lines |
| *INCENP* | 1.00 | PC18 | ENST00000278849.4: c.2702delC | key regulator of mitosis, controls the kinetochore localisation of BUB1 |
| *NAP1L4* | 1.00 | PC21 | ENST00000380542.4: c.20C>G | possible histone chaperone |
| *PRDM2* | 1.00 | PC53 | ENST00000413440.1: c.249delT | S-adenosyl-L-methionine-dependent histone methyltransferase that specifically methylates 'Lys-9' of histone H3 |
| *PRMT5* | 0.99 | PC16 | ENST00000324366.8: c.799_800insA | post-translational arginine methylation, suspected oncogene |
| *LMTK2* | 0.99 | PC53 | ENST00000297293.5: c.149_150delTG | phosphorylates PPP1C |

pLI score rounded up to 2 decimal places.

**Table A-2 (continued). Thirty genes with an ExAC pLI score>0.9 and a LoF variant in the discovery cohort**

| Gene | pLI | Sample | Variant | Gene function from GeneCards or PubMed |
|---|---|---|---|---|
| *SIN3B* | 0.99 | PC35 | ENST00000379803.1: c.784_785insC | transcriptional repressor, interacts with MXI1 to repress Myc-responsive genes |
| *JAG2* | 0.99 | PC36 | ENST00000347004.2: c.3314_3315insG | Putative Notch ligand involved in the mediation of Notch signaling |
| *SREK1* | 0.99 | PC44 | ENST00000380918.3: c.1253_1256delAGAG | regulates alternative splicing |
| *CACNA1B* | 0.98 | PC38 | ENST00000371372.1: c.390+1_390+2insACGACACGGAGCCCTATTTCATCGGGATCTTTTGCTTCGAGGCAGGGA | calcium channel |
| *GRM4* | 0.98 | PC13 | ENST00000374177.3: c.310C>T | glutamate receptor in the brain |
| *ZC3H3* | 0.96 | PC32 | ENST00000262577.5: c.2811_2815+49delAGGAACCACTCCCCCCGGGACCCCCGGGCTCGGTCCGGCACGACGGATGGACTC | Required for the export of polyadenylated mRNAs from the nucleus |
| *CBX2* | 0.95 | PC14 | ENST00000269399.5: c.460delT | Component of a Polycomb group (PcG) multiprotein PRC1-like complex, a complex class required to maintain the transcriptionally repressive state of many genes, including Hox genes, throughout development |
| *PHRF1* | 0.95 | PC49 | ENST00000264555.5: c.2944_2945delCA | ubiquitin ligase, postulated tumour suppressor gene through interactions with TGFB signalling |
| *AKAP10* | 0.93 | PC14 | ENST00000225737.6: c.37_38insC | mitochondrial protein |
| *MAP7* | 0.93 | PC4 | ENST00000544465.1: c.122-1G>C | microtubule stabilising protein |
| *IQSEC3* | 0.93 | PC18 | ENST00000326261.4: c.2806_2807insAGTGTAAGTCTTTGACAGCC | guanine nucleotide exchange factor |

pLI score rounded up to 2 decimal places.

# List of References


AALTONEN, L., PELTOMAKI, P., LEACH, F., SISTONEN, P., PYLKKANEN, L., MECKLIN, J., JARVINEN, H., POWELL, S., JEN, J., HAMILTON & ET, A. 1993. Clues to the pathogenesis of familial colorectal cancer. *Science,* 260**,** 812-816.

ACHARYA, S., WILSON, T., GRADIA, S., KANE, M. F., GUERRETTE, S., MARSISCHKY, G. T., KOLODNER, R. & FISHEL, R. 1996. hMSH2 forms specific mispair-binding complexes with hMSH3 and hMSH6. *Proc Natl Acad Sci U S A,* 93**,** 13629-34.

ADAIMY, L., CHOUERY, E., MÉGARBANÉ, H., MROUEH, S., DELAGUE, V., NICOLAS, E., BELGUITH, H., DE MAZANCOURT, P. & MÉGARBANÉ, A. 2007. Mutation in WNT10A is associated with an autosomal recessive ectodermal dysplasia: The odonto-onycho-dermal dysplasia. *American Journal of Human Genetics,* 81**,** 821-828.

AKIYAMA, Y., SATO, H., YAMADA, T., NAGASAKI, H., TSUCHIYA, A., ABE, R. & YUASA, Y. 1997. Germ-line mutation of the hMSH6/GTBP gene in an atypical hereditary nonpolyposis colorectal cancer kindred. *Cancer Res,* 57**,** 3920-3.

AL-SALEEM, T., WESSNER, L. L., SCHEITHAUER, B. W., PATTERSON, K., ROACH, E. S., DREYER, S. J., FUJIKAWA, K., BJORNSSON, J., BERNSTEIN, J. & HENSKE, E. P. 1998. Malignant tumors of the kidney, brain, and soft tissues in children and young adults with the tuberous sclerosis complex. *Cancer,* 83**,** 2208-2216.

AL-TASSAN, N., CHMIEL, N. H., MAYNARD, J., FLEMING, N., LIVINGSTON, A. L., WILLIAMS, G. T., HODGES, A. K., DAVIES, D. R., DAVID, S. S., SAMPSON, J. R. & CHEADLE, J. P. 2002. Inherited variants of MYH associated with somatic G:C→T:A mutations in colorectal tumors. *Nature Genetics,* 30**,** 227-232.

AL-TASSAN, N. A., WHIFFIN, N., HOSKING, F. J., PALLES, C., FARRINGTON, S. M., DOBBINS, S. E., HARRIS, R., GORMAN, M., TENESA, A., MEYER, B. F., WAKIL, S. M., KINNERSLEY, B., CAMPBELL, H., MARTIN, L., SMITH, C. G., IDZIASZCZYK, S., BARCLAY, E., MAUGHAN, T. S., KAPLAN, R., KERR, R., KERR, D., BUCHANAN, D. D., WIN, A. K., HOPPER, J., JENKINS, M., LINDOR, N. M., NEWCOMB, P. A., GALLINGER, S., CONTI, D., SCHUMACHER, F., CASEY, G., DUNLOP, M. G., TOMLINSON, I. P., CHEADLE, J. P. & HOULSTON, R. S. 2015. A new GWAS and meta-analysis with 1000Genomes imputation identifies novel risk variants for colorectal cancer. *Sci Rep,* 5**,** 10442.

ALEXANDROV, L. B., NIK-ZAINAL, S., WEDGE, D. C., APARICIO, S. A. J. R., BEHJATI, S., BIANKIN, A. V., BIGNELL, G. R., BOLLI, N., BORG, A., BØRRESEN-DALE, A.-L., BOYAULT, S., BURKHARDT, B., BUTLER, A. P., CALDAS, C., DAVIES, H. R., DESMEDT, C., EILS, R., EYFJÖRD, J. E., FOEKENS, J. A., GREAVES, M., HOSODA, F., HUTTER, B., ILICIC, T., IMBEAUD, S., IMIELINSKI, M., JÄGER, N., JONES, D. T. W., JONES, D., KNAPPSKOG, S., KOOL, M., LAKHANI, S. R., LÓPEZ-OTÍN, C., MARTIN, S., MUNSHI, N. C., NAKAMURA, H., NORTHCOTT, P. A., PAJIC, M., PAPAEMMANUIL, E., PARADISO, A., PEARSON, J. V., PUENTE, X. S., RAINE, K., RAMAKRISHNA, M., RICHARDSON, A. L., RICHTER, J., ROSENSTIEL, P., SCHLESNER, M., SCHUMACHER, T. N., SPAN, P. N., TEAGUE, J. W., TOTOKI, Y., TUTT, A. N. J., VALDÉS-MAS, R., VAN BUUREN, M. M., VAN 'T VEER, L., VINCENT-SALOMON, A., WADDELL, N., YATES, L. R., AUSTRALIAN PANCREATIC CANCER GENOME, I., CONSORTIUM, I. B. C., CONSORTIUM, I. M.-S., PEDBRAIN, I., ZUCMAN-ROSSI, J., ANDREW FUTREAL, P., MCDERMOTT, U., LICHTER, P., MEYERSON, M., GRIMMOND, S. M., SIEBERT, R., CAMPO, E., SHIBATA, T., PFISTER, S. M., CAMPBELL, P. J. & STRATTON, M. R. 2013. Signatures of mutational processes in human cancer. *Nature,* 500**,** 415.

ALTSHULER, D., DONNELLY, P. & THE INTERNATIONAL HAPMAP, C. 2005. A haplotype map of the human genome. *Nature,* 437**,** 1299-1320.

ARETZ, S., STIENEN, D., UHLHAAS, S., STOLTE, M., ENTIUS, M. M., LOFF, S., BACK, W., KAUFMANN, A., KELLER, K. M., BLAAS, S. H., SIEBERT, R., VOGT, S., SPRANGER, S.,

HOLINSKI-FEDER, E., SUNDE, L., PROPPING, P. & FRIEDL, W. 2007. High proportion of large genomic deletions and a genotype phenotype update in 80 unrelated families with juvenile polyposis syndrome. *J Med Genet,* 44**,** 702-9.

ARORA, S., YAN, H., CHO, I., FAN, H. Y., LUO, B., GAI, X., BODIAN, D. L., VOCKLEY, J. G., ZHOU, Y., HANDORF, E. A., EGLESTON, B. L., ANDRAKE, M. D., NICOLAS, E., SEREBRIISKII, I. G., YEN, T. J., HALL, M. J., GOLEMIS, E. A. & ENDERS, G. H. 2015. Genetic Variants That Predispose to DNA Double-Strand Breaks in Lymphocytes From a Subset of Patients With Familial Colorectal Carcinomas. *Gastroenterology,* 149**,** 1872-1883 e9.

ARTAVANIS-TSAKONAS, S., RAND, M. D. & LAKE, R. J. 1999. Notch signaling: cell fate control and signal integration in development. *Science,* 284**,** 770-6.

ASHBURNER, M., BALL, C. A., BLAKE, J. A., BOTSTEIN, D., BUTLER, H., CHERRY, J. M., DAVIS, A. P., DOLINSKI, K., DWIGHT, S. S., EPPIG, J. T., HARRIS, M. A., HILL, D. P., ISSEL-TARVER, L., KASARSKIS, A., LEWIS, S., MATESE, J. C., RICHARDSON, J. E., RINGWALD, M., RUBIN, G. M. & SHERLOCK, G. 2000. Gene Ontology: tool for the unification of biology. *Nat Genet,* 25**,** 25-29.

AUSTRALIAN INSTITUTE OF HEALTH AND WELFARE 2017. *Cancer In Australia 2017,* Canberra, Australian Institute of Health and Welfare.

BACHMANN-GAGESCU, R., DEMPSEY, J. C., PHELPS, I. G., O'ROAK, B. J., KNUTZEN, D. M., RUE, T. C., ISHAK, G. E., ISABELLA, C. R., GORDEN, N., ADKINS, J., BOYLE, E. A., DE LACY, N., O'DAY, D., ALSWAID, A., RAMADEVI A, R., LINGAPPA, L., LOURENÇO, C., MARTORELL, L., GARCIA-CAZORLA, À., OZYÜREK, H., HALILOĞLU, G., TUYSUZ, B., TOPÇU, M., CHANCE, P., PARISI, M. A., GLASS, I. A., SHENDURE, J. & DOHERTY, D. 2015. Joubert syndrome: a model for untangling recessive disorders with extreme genetic heterogeneity. *Journal of Medical Genetics,* 52**,** 514-522.

BALAGUER, F., CASTELLVÍ–BEL, S., CASTELLS, A., ANDREU, M., MUÑOZ, J., GISBERT, J. P., LLOR, X., JOVER, R., DE CID, R., GONZALO, V., BESSA, X., XICOLA, R. M., PONS, E., ALENDA, C., PAYÁ, A. & PIQUÉ, J. M. 2007. Identification of MYH Mutation Carriers in Colorectal Cancer: A Multicenter, Case-Control, Population-Based Study. *Clinical Gastroenterology and Hepatology,* 5**,** 379-387.

BAMSHAD, M. J., NG, S. B., BIGHAM, A. W., TABOR, H. K., EMOND, M. J., NICKERSON, D. A. & SHENDURE, J. 2011. Exome sequencing as a tool for Mendelian disease gene discovery. *Nature Reviews Genetics,* 12**,** 745.

BARTSCH, O., SCHMIDT, S., RICHTER, M., MORLOT, S., SEEMANOVA, E., WIEBE, G. & RASI, S. 2005. DNA sequencing of CREBBP demonstrates mutations in 56% of patients with Rubinstein-Taybi syndrome (RSTS) and in another patient with incomplete RSTS. *Hum Genet,* 117**,** 485-93.

BELHADJ, S., QUINTANA, I., MUR, P., MUNOZ-TORRES, P. M., ALONSO, M. H., NAVARRO, M., TERRADAS, M., PIÑOL, V., BRUNET, J., MORENO, V., LÁZARO, C., CAPELLÁ, G. & VALLE, L. 2019. NTHL1 biallelic mutations seldom cause colorectal cancer, serrated polyposis or a multi-tumor phenotype, in absence of colorectal adenomas. *Scientific Reports,* 9**,** 9020.

BELLIDO, F., PINEDA, M., AIZA, G., VALDÉS-MAS, R., NAVARRO, M., PUENTE, D. A., PONS, T., GONZÁLEZ, S., IGLESIAS, S., DARDER, E., PIÑOL, V., SOTO, J. L., VALENCIA, A., BLANCO, I., URIOSTE, M., BRUNET, J., LÁZARO, C., CAPELLÁ, G., PUENTE, X. S. & VALLE, L. 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. *Genetics in Medicine,* 18**,** 325-332.

BENTLEY, D. R., BALASUBRAMANIAN, S., SWERDLOW, H. P., SMITH, G. P., MILTON, J., BROWN, C. G., HALL, K. P., EVERS, D. J., BARNES, C. L., BIGNELL, H. R., BOUTELL, J. M., BRYANT, J., CARTER, R. J., KEIRA CHEETHAM, R., COX, A. J., ELLIS, D. J., FLATBUSH, M. R., GORMLEY, N. A., HUMPHRAY, S. J., IRVING, L. J., KARBELASHVILI, M. S., KIRK, S. M., LI,

H., LIU, X., MAISINGER, K. S., MURRAY, L. J., OBRADOVIC, B., OST, T., PARKINSON, M. L., PRATT, M. R., RASOLONJATOVO, I. M. J., REED, M. T., RIGATTI, R., RODIGHIERO, C., ROSS, M. T., SABOT, A., SANKAR, S. V., SCALLY, A., SCHROTH, G. P., SMITH, M. E., SMITH, V. P., SPIRIDOU, A., TORRANCE, P. E., TZONEV, S. S., VERMAAS, E. H., WALTER, K., WU, X., ZHANG, L., ALAM, M. D., ANASTASI, C., ANIEBO, I. C., BAILEY, D. M. D., BANCARZ, I. R., BANERJEE, S., BARBOUR, S. G., BAYBAYAN, P. A., BENOIT, V. A., BENSON, K. F., BEVIS, C., BLACK, P. J., BOODHUN, A., BRENNAN, J. S., BRIDGHAM, J. A., BROWN, R. C., BROWN, A. A., BUERMANN, D. H., BUNDU, A. A., BURROWS, J. C., CARTER, N. P., CASTILLO, N., CHIARA E. CATENAZZI, M., CHANG, S., NEIL COOLEY, R., CRAKE, N. R., DADA, O. O., DIAKOUMAKOS, K. D., DOMINGUEZ-FERNANDEZ, B., EARNSHAW, D. J., EGBUJOR, U. C., ELMORE, D. W., ETCHIN, S. S., EWAN, M. R., FEDURCO, M., FRASER, L. J., FUENTES FAJARDO, K. V., SCOTT FUREY, W., GEORGE, D., GIETZEN, K. J., GODDARD, C. P., GOLDA, G. S., GRANIERI, P. A., GREEN, D. E., GUSTAFSON, D. L., HANSEN, N. F., HARNISH, K., HAUDENSCHILD, C. D., HEYER, N. I., HIMS, M. M., HO, J. T., HORGAN, A. M., et al. 2008. Accurate whole human genome sequencing using reversible terminator chemistry. *Nature,* 456**,** 53.

BERTARIO, L., RUSSO, A., SALA, P., EBOLI, M., GIAROLA, M., D'AMICO, F., GISMONDI, V., VARESCO, L., PIEROTTI, M. A., RADICE, P. & ON BEHALF OF THE HEREDITARY COLORECTAL TUMOURS, R. 2001. Genotype and phenotype factors as determinants of desmoid tumors in patients with familial adenomatous polyposis. *International Journal of Cancer,* 95**,** 102-107.

BERTARIO, L., RUSSO, A., SALA, P., VARESCO, L., GIAROLA, M., MONDINI, P., PIEROTTI, M., SPINELLI, P. & RADICE, P. 2003. Multiple Approach to the Exploration of Genotype-Phenotype Correlations in Familial Adenomatous Polyposis. *Journal of Clinical Oncology,* 21**,** 1698-1707.

BETTINGTON, M., WALKER, N., CLOUSTON, A., BROWN, I., LEGGETT, B. & WHITEHALL, V. 2013. The serrated pathway to colorectal carcinoma: current concepts and challenges. *Histopathology,* 62**,** 367-86.

BISGAARD, M. L., FENGER, K., BÜLOW, S., NIEBUHR, E. & MOHR, J. 1994. Familial adenomatous polyposis (FAP): Frequency, penetrance, and mutation rate. *Human Mutation,* 3**,** 121-125.

BJÖRK, J. 1999. Epidemiology of Familial Adenomatous Polyposis in Sweden: Changes over Time and Differences in Phenotype between Males and Females. *Scandinavian Journal of Gastroenterology,* 34**,** 1230-1235.

BLACKWELL, L. J., WANG, S. & MODRICH, P. 2001. DNA chain length dependence of formation and dynamics of hMutSalpha.hMutLalpha.heteroduplex complexes. *J Biol Chem,* 276**,** 33233-40.

BODMER, W. F., BAILEY, C. J., BODMER, J., BUSSEY, H. J., ELLIS, A., GORMAN, P., LUCIBELLO, F. C., MURDAY, V. A., RIDER, S. H., SCAMBLER, P. & ET AL. 1987. Localization of the gene for familial adenomatous polyposis on chromosome 5. *Nature,* 328**,** 614-6.

BONADONA, V., BONAÏTI, B., OLSCHWANG, S. & ET AL. 2011. Cancer risks associated with germline mutations in mlh1, msh2, and msh6 genes in lynch syndrome. *JAMA,* 305**,** 2304-2310.

BOSMAN, F., CARNEIRO, F., HRUBAN, R. & THEISE, N. E. 2010. *WHO Classification of Tumours of the Digestive System. ,* Berlin, Springer-Verlag.

BOTSTEIN, D. & RISCH, N. 2003. Discovering genotypes underlying human phenotypes: past successes for mendelian disease, future approaches for complex disease. *Nature Genetics,* 33 Suppl**,** 228-237.

BOURDEAUT, F., MIQUEL, C., RICHER, W., GRILL, J., ZERAH, M., GRISON, C., PIERRON, G., AMIEL, J., KRUCKER, C., RADVANYI, F., BRUGIERES, L. & DELATTRE, O. 2014. Rubinstein-

Taybi syndrome predisposing to non-WNT, non-SHH, group 3 medulloblastoma. *Pediatric Blood and Cancer,* 61**,** 383-386.

BOURSI, B., SELLA, T., LIBERMAN, E., SHAPIRA, S., DAVID, M., KAZANOV, D., ARBER, N. & KRAUS, S. 2013. The APC p.I1307K polymorphism is a significant risk factor for CRC in average risk Ashkenazi Jews. *European Journal of Cancer,* 49**,** 3680-3685.

BOYCE, S., NASSAR, N., LEE, C. Y., SUEN, M. K., AL ZAHRANI, S. & GLADMAN, M. A. 2016. Young-onset colorectal cancer in New South Wales: a population-based study. *Med J Aust,* 205**,** 465-470.

BRIGGS, S. & TOMLINSON, I. 2013. Germline and somatic polymerase ε and δ mutations define a new class of hypermutated colorectal and endometrial cancers. *The Journal of Pathology,* 230**,** 148-153.

BRINKMAN, J., DE NEF, J. J., BARTH, P. G. & VERSCHUUR, A. C. 2005. Burkitt lymphoma in a child with Joubert syndrome. *Pediatr Blood Cancer,* 44**,** 397-9.

BROAD INSTITUTE. 2016. *GATK Understanding and adapting the generic hard-filtering recommendations* [Online]. Available: https://software.broadinstitute.org/gatk/documentation/article?id=6925 [Accessed].

BRODERICK, P., CARVAJAL-CARMONA, L., PITTMAN, A. M., WEBB, E., HOWARTH, K., ROWAN, A., LUBBE, S., SPAIN, S., SULLIVAN, K., FIELDING, S., JAEGER, E., VIJAYAKRISHNAN, J., KEMP, Z., GORMAN, M., CHANDLER, I., PAPAEMMANUIL, E., PENEGAR, S., WOOD, W., SELLICK, G., QURESHI, M., TEIXEIRA, A., DOMINGO, E., BARCLAY, E., MARTIN, L., SIEBER, O., KERR, D., GRAY, R., PETO, J., CAZIER, J. B., TOMLINSON, I. & HOULSTON, R. S. 2007. A genome-wide association study shows that common alleles of SMAD7 influence colorectal cancer risk. *Nat Genet,* 39**,** 1315-7.

BRODERICK, P., DOBBINS, S. E., CHUBB, D., KINNERSLEY, B., DUNLOP, M. G., TOMLINSON, I. & HOULSTON, R. S. 2016. Validation of Recently Proposed Colorectal Cancer Susceptibility Gene Variants in an Analysis of Families and Patients-a Systematic Review. *Gastroenterology*.

BRONNER, C. E., BAKER, S. M., MORRISON, P. T., WARREN, G., SMITH, L. G., LESCOE, M. K., KANE, M., EARABINO, C., LIPFORD, J., LINDBLOM, A., TANNERGARD, P., BOLLAG, R. J., GODWIN, A. R., WARD, D. C., NORDENSKJLD, M., FISHEL, R., KOLODNER, R. & LISKAY, R. M. 1994. Mutation in the DNA mismatch repair gene homologue hMLH 1 is associated with hereditary non-polyposis colon cancer. *Nature,* 368**,** 258-261.

BUCHANAN, D. D., CLENDENNING, M., ZHUOER, L., STEWART, J. R., JOSELAND, S., WOODALL, S., ARNOLD, J., SEMOTIUK, K., ARONSON, M., HOLTER, S., GALLINGER, S., JENKINS, M. A., SWEET, K., MACRAE, F. A., WINSHIP, I. M., PARRY, S. & ROSTY, C. 2016. Lack of evidence for germline RNF43 mutations in patients with serrated polyposis syndrome from a large multinational study. *Gut*.

BUTLER, G. H., BOYLE, M., LYNCH, S. A., RYAN, S., MCDERMOTT, M. & CAPRA, M. 2016. One to watch: A germ cell tumor arising in an undescended testicle in Rubinstein-Taybi syndrome. *Journal of Pediatric Hematology/Oncology,* 38**,** e191-e192.

CARBON, S., IRELAND, A., MUNGALL, C. J., SHU, S., MARSHALL, B. & LEWIS, S. 2009. AmiGO: online access to ontology and annotation data. *Bioinformatics,* 25**,** 288-289.

CARMENA, M., WHEELOCK, M., FUNABIKI, H. & EARNSHAW, W. C. 2012. The chromosomal passenger complex (CPC): From easy rider to the godfather of mitosis. *Nature Reviews Molecular Cell Biology,* 13**,** 789-803.

CARVAJAL-CARMONA, L. G., CAZIER, J.-B., JONES, A. M., HOWARTH, K., BRODERICK, P., PITTMAN, A., DOBBINS, S., TENESA, A., FARRINGTON, S., PRENDERGAST, J., THEODORATOU, E., BARNETSON, R., CONTI, D., NEWCOMB, P., HOPPER, J. L., JENKINS, M. A., GALLINGER, S., DUGGAN, D. J., CAMPBELL, H., KERR, D., CASEY, G., HOULSTON, R., DUNLOP, M. & TOMLINSON, I. 2011. Fine-mapping of colorectal cancer susceptibility loci at 8q23.3, 16q22.1 and 19q13.11: refinement of association signals

and use of in silico analysis to suggest functional variation and unexpected candidate target genes. *Human Molecular Genetics,* 20**,** 2879-2888.

CARVAJAL CARMONA, L. G. & TOMLINSON, I. 2016. The Hunting of the Snark: Whither Genome-Wide Association Studies for Colorectal Cancer? *Gastroenterology,* 150**,** 1528-30.

CASCALHO, M., WONG, J., STEINBERG, C. & WABL, M. 1998. Mismatch Repair Co-opted by Hypermutation. *Science,* 279**,** 1207-1210.

CHANG, Y. F., IMAM, J. S. & WILKINSON, M. F. 2007. The Nonsense-mediated decay RNA surveillance pathway. *Annual Review of Biochemistry.*

CHEADLE, J. P. & SAMPSON, J. R. 2007. MUTYH-associated polyposis—From defect in base excision repair to clinical genetic testing. *DNA Repair,* 6**,** 274-279.

CHOW, E., LIPTON, L., LYNCH, E., D'SOUZA, R., ARAGONA, C., HODGKIN, L., BROWN, G., WINSHIP, I., BARKER, M., BUCHANAN, D., COWIE, S., NASIOULAS, S., DU SART, D., YOUNG, J., LEGGETT, B., JASS, J. & MACRAE, F. 2006. Hyperplastic polyposis syndrome: phenotypic presentations and the role of MBD4 and MYH. *Gastroenterology,* 131**,** 30-9.

CHUBB, D., BRODERICK, P., DOBBINS, S. E., FRAMPTON, M., KINNERSLEY, B., PENEGAR, S., PRICE, A., MA, Y. P., SHERBORNE, A. L., PALLES, C., TIMOFEEVA, M. N., BISHOP, D. T., DUNLOP, M. G., TOMLINSON, I. & HOULSTON, R. S. 2016. Rare disruptive mutations and their contribution to the heritable risk of colorectal cancer. *Nat Commun,* 7**,** 11883.

CHUBB, D., BRODERICK, P., FRAMPTON, M., KINNERSLEY, B., SHERBORNE, A., PENEGAR, S., LLOYD, A., MA, Y. P., DOBBINS, S. E. & HOULSTON, R. S. 2015. Genetic diagnosis of high-penetrance susceptibility for colorectal cancer (CRC) is achievable for a high proportion of familial CRC by exome sequencing. *J Clin Oncol,* 33**,** 426-32.

CLEARY, S. P., COTTERCHIO, M., JENKINS, M. A., KIM, H., BRISTOW, R., GREEN, R., HAILE, R., HOPPER, J. L., LEMARCHAND, L., LINDOR, N., PARFREY, P., POTTER, J., YOUNGHUSBAND, B. & GALLINGER, S. 2009. Germline MutY Human Homologue Mutations and Colorectal Cancer: A Multisite Case-Control Study. *Gastroenterology,* 136**,** 1251-1260.

CLENDENNING, M., HAMPEL, H., LAJEUNESSE, J., LINDBLOM, A., LOCKMAN, J., NILBERT, M., SENTER, L., SOTAMAA, K. & DE LA CHAPELLE, A. 2006. Long-range PCR facilitates the identification of PMS2-specific mutations. *Hum Mutat,* 27**,** 490-5.

CLENDENNING, M., SENTER, L., HAMPEL, H., ROBINSON, K. L., SUN, S., BUCHANAN, D., WALSH, M. D., NILBERT, M., GREEN, J., POTTER, J., LINDBLOM, A. & DE LA CHAPELLE, A. 2008. A frame-shift mutation of PMS2 is a widespread cause of Lynch syndrome. *J Med Genet,* 45**,** 340-5.

CLENDENNING, M., YOUNG, J. P., WALSH, M. D., WOODALL, S., ARNOLD, J., JENKINS, M., WIN, A. K., HOPPER, J. L., SWEET, K., GALLINGER, S., ROSTY, C., PARRY, S. & BUCHANAN, D. D. 2013. Germline Mutations in the Polyposis-Associated Genes BMPR1A, SMAD4, PTEN, MUTYH and GREM1 Are Not Common in Individuals with Serrated Polyposis Syndrome. *PLoS ONE,* 8.

COMINO-MÉNDEZ, I., GRACIA-AZNÁREZ, F. J., SCHIAVI, F., LANDA, I., LEANDRO-GARCÍA, L. J., LETÓN, R., HONRADO, E., RAMOS-MEDINA, R., CARONIA, D., PITA, G., GÓMEZ-GRAÑA, A., DE CUBAS, A. A., INGLADA-PÉREZ, L., MALISZEWSKA, A., TASCHIN, E., BOBISSE, S., PICA, G., LOLI, P., HERNÁNDEZ-LAVADO, R., DÍAZ, J. A., GÓMEZ-MORALES, M., GONZÁLEZ-NEIRA, A., RONCADOR, G., RODRÍGUEZ-ANTONA, C., BENÍTEZ, J., MANNELLI, M., OPOCHER, G., ROBLEDO, M. & CASCÓN, A. 2011. Exome sequencing identifies MAX mutations as a cause of hereditary pheochromocytoma. *Nature Genetics,* 43**,** 663-667.

COOPER, G. M., STONE, E. A., ASIMENOS, G., GREEN, E. D., BATZOGLOU, S. & SIDOW, A. 2005. Distribution and intensity of constraint in mammalian genomic sequence. *Genome Res,* 15**,** 901-13.

CROWN HUMAN GENOME CENTER. 1997. *GeneCards® Human Gene Database* [Online]. Available: https://www.genecards.org/ [Accessed 16/12/2015].

CUI, R., OKADA, Y., JANG, S. G., KU, J. L., PARK, J. G., KAMATANI, Y., HOSONO, N., TSUNODA, T., KUMAR, V., TANIKAWA, C., KAMATANI, N., YAMADA, R., KUBO, M., NAKAMURA, Y. & MATSUDA, K. 2011. Common variant in 6q26-q27 is associated with distal colon cancer in an Asian population. *Gut,* 60**,** 799-805.

DE JONG, A. E., HENDRIKS, Y. M. C., KLEIBEUKER, J. H., DE BOER, S. Y., CATS, A., GRIFFIOEN, G., NAGENGAST, F. M., NELIS, F. G., ROOKUS, M. A. & VASEN, H. F. A. 2006. Decrease in mortality in Lynch syndrome families because of surveillance. *Gastroenterology,* 130**,** 665-671.

DE VOER, R. M., GEURTS VAN KESSEL, A., WEREN, R. D., LIGTENBERG, M. J., SMEETS, D., FU, L., VREEDE, L., KAMPING, E. J., VERWIEL, E. T., HAHN, M. M., ARIAANS, M., SPRUIJT, L., VAN ESSEN, T., HOUGE, G., SCHACKERT, H. K., SHENG, J. Q., VENSELAAR, H., VAN RAVENSWAAIJ-ARTS, C. M., VAN KRIEKEN, J. H., HOOGERBRUGGE, N. & KUIPER, R. P. 2013. Germline mutations in the spindle assembly checkpoint genes BUB1 and BUB3 are risk factors for colorectal cancer. *Gastroenterology,* 145**,** 544-7.

DEGROLARD-COURCET, E., SOKOLOWSKA, J., PADEANO, M. M., GUIU, S., BRONNER, M., CHERY, C., CORON, F., LEPAGE, C., CHAPUSOT, C., LOUSTALOT, C., JOUVE, J. L., HATEM, C., FERRANT, E., MARTIN, L., COUTANT, C., BAURAND, A., COUILLAULT, G., DELIGNETTE, A., EL CHEHADEH, S., LIZARD, S., ARNOULD, L., FUMOLEAU, P., CALLIER, P., MUGNERET, F., PHILIPPE, C., FREBOURG, T., JONVEAUX, P. & FAIVRE, L. 2014. Development of primary early-onset colorectal cancers due to biallelic mutations of the FANCD1/BRCA2 gene. *Eur J Hum Genet,* 22**,** 979-87.

DEMPSEY, J. C., PHELPS, I. G., BACHMANN-GAGESCU, R., GLASS, I. A., TULLY, H. M. & DOHERTY, D. 2017. Mortality in Joubert syndrome. *Am J Med Genet A,* 173**,** 1237-1242.

DEN DUNNEN, J. T. & VAN OMMEN, G.-J. B. 1999. The protein truncation test: A review. *Human Mutation,* 14**,** 95-102.

DERYCKE, M. S., GUNAWARDENA, S. R., MIDDHA, S., ASMANN, Y. W., SCHAID, D. J., MCDONNELL, S. K., RISKA, S. M., ECKLOFF, B. W., CUNNINGHAM, J. M., FRIDLEY, B. L., SERIE, D. J., BAMLET, W. R., CICEK, M. S., JENKINS, M. A., DUGGAN, D. J., BUCHANAN, D., CLENDENNING, M., HAILE, R. W., WOODS, M. O., GALLINGER, S. N., CASEY, G., POTTER, J. D., NEWCOMB, P. A., LE MARCHAND, L., LINDOR, N. M., THIBODEAU, S. N. & GOODE, E. L. 2013. Identification of Novel Variants in Colorectal Cancer Families by High-Throughput Exome Sequencing. *Cancer Epidemiol Biomarkers Prev*.

DESAI, D. C., LOCKMAN, J. C., CHADWICK, R. B., GAO, X., PERCESEPE, A., EVANS, D. G. R., MIYAKI, M., YUEN, S. T., RADICE, P., MAHER, E. R., WRIGHT, F. A. & DE LA CHAPELLE, A. 2000. Recurrent germline mutation in *MSH2* arises frequently de novo. *Journal of Medical Genetics,* 37**,** 646-652.

DESAI, D. C., MURDAY, V., PHILLIPS, R. K., NEALE, K. F., MILLA, P. & HODGSON, S. V. 1998. A survey of phenotypic features in juvenile polyposis. *J Med Genet,* 35**,** 476-81.

DOBBINS, S. E., BRODERICK, P., CHUBB, D., KINNERSLEY, B., SHERBORNE, A. L. & HOULSTON, R. S. 2016. Undefined familial colorectal cancer and the role of pleiotropism in cancer susceptibility genes. *Familial Cancer,* 15**,** 593-599.

DOMINGO, E., LAIHO, P., OLLIKAINEN, M., PINTO, M., WANG, L., FRENCH, A. J., WESTRA, J., FREBOURG, T., ESPÍN, E., ARMENGOL, M., HAMELIN, R., YAMAMOTO, H., HOFSTRA, R. M. W., SERUCA, R., LINDBLOM, A., PELTOMÄKI, P., THIBODEAU, S. N., AALTONEN, L. A. & SCHWARTZ, S. 2004. BRAF screening as a low-cost effective strategy for simplifying HNPCC genetic testing. *Journal of Medical Genetics,* 41**,** 664-668.

DOZOIS, E. J., BOARDMAN, L. A., SUWANTHANMA, W., LIMBURG, P. J., CIMA, R. R., BAKKEN, J. L., VIERKANT, R. A., AAKRE, J. A. & LARSON, D. W. 2008. Young-Onset Colorectal Cancer in Patients With No Known Genetic Predisposition: Can We Increase Early Recognition and Improve Outcome? *Medicine,* 87**,** 259-263.

DUNLOP, M. G., DOBBINS, S. E., FARRINGTON, S. M., JONES, A. M., PALLES, C., WHIFFIN, N., TENESA, A., SPAIN, S., BRODERICK, P., OOI, L. Y., DOMINGO, E., SMILLIE, C., HENRION, M., FRAMPTON, M., MARTIN, L., GRIMES, G., GORMAN, M., SEMPLE, C., MA, Y. P., BARCLAY, E., PRENDERGAST, J., CAZIER, J. B., OLVER, B., PENEGAR, S., LUBBE, S., CHANDER, I., CARVAJAL-CARMONA, L. G., BALLEREAU, S., LLOYD, A., VIJAYAKRISHNAN, J., ZGAGA, L., RUDAN, I., THEODORATOU, E., STARR, J. M., DEARY, I., KIRAC, I., KOVACEVIC, D., AALTONEN, L. A., RENKONEN-SINISALO, L., MECKLIN, J. P., MATSUDA, K., NAKAMURA, Y., OKADA, Y., GALLINGER, S., DUGGAN, D. J., CONTI, D., NEWCOMB, P., HOPPER, J., JENKINS, M. A., SCHUMACHER, F., CASEY, G., EASTON, D., SHAH, M., PHAROAH, P., LINDBLOM, A., LIU, T., SMITH, C. G., WEST, H., CHEADLE, J. P., MIDGLEY, R., KERR, D. J., CAMPBELL, H., TOMLINSON, I. P. & HOULSTON, R. S. 2012. Common variation near CDKN1A, POLD3 and SHROOM2 influences colorectal cancer risk. *Nat Genet,* 44**,** 770-6.

EIDE, P. W., CEKAITE, L., DANIELSEN, S. A., EILERTSEN, I. A., KJENSETH, A., FYKERUD, T. A., AGESEN, T. H., BRUUN, J., RIVEDAL, E., LOTHE, R. A. & LEITHE, E. 2013. NEDD4 is overexpressed in colorectal cancer and promotes colonic cell growth independently of the PI3K/PTEN/AKT pathway. *Cell Signal,* 25**,** 12-8.

ENGEL, C., LOEFFLER, M., STEINKE, V., RAHNER, N., HOLINSKI-FEDER, E., DIETMAIER, W., SCHACKERT, H. K., GOERGENS, H., VON KNEBEL DOEBERITZ, M., GOECKE, T. O., SCHMIEGEL, W., BUETTNER, R., MOESLEIN, G., LETTEBOER, T. G., GOMEZ GARCIA, E., HES, F. J., HOOGERBRUGGE, N., MENKO, F. H., VAN OS, T. A., SIJMONS, R. H., WAGNER, A., KLUIJT, I., PROPPING, P. & VASEN, H. F. 2012. Risks of less common cancers in proven mutation carriers with lynch syndrome. *J Clin Oncol,* 30**,** 4409-15.

ESTEBAN-JURADO, C., VILA-CASADESÚS, M., GARRE, P., LOZANO, J. J., PRISTOUPILOVA, A., BELTRAN, S., MUÑOZ, J., OCAÑA, T., BALAGUER, F., LÓPEZ-CERÓN, M., CUATRECASAS, M., FRANCH-EXPÓSITO, S., PIQUÉ, J. M., CASTELLS, A., CARRACEDO, A., RUIZ-PONTE, C., ABULÍ, A., BESSA, X., ANDREU, M., BUJANDA, L., CALDÉS, T. & CASTELLVÍ-BEL, S. 2014. Whole-exome sequencing identifies rare pathogenic variants in new predisposition genes for familial colorectal cancer. *Genetics In Medicine,* 17**,** 131.

ETTAHAR, A., FERRIGNO, O., ZHANG, M.-Z., OHNISHI, M., FERRAND, N., PRUNIER, C., LEVY, L., BOURGEADE, M.-F., BIECHE, I., ROMERO, DAMIAN G., COLLAND, F. & ATFI, A. 2013. Identification of PHRF1 as a Tumor Suppressor that Promotes the TGF-β Cytostatic Program through Selective Release of TGIF-Driven PML Inactivation. *Cell Reports,* 4**,** 530-541.

EXOME VARIANT SERVER. n.d. *NHLBI GO Exome Sequencing Project (ESP), Seattle, WA (URL: http://evs.gs.washington.edu/EVS/)* [Online]. [Accessed].

FAN, J., WEI, Q., LIAO, J., ZOU, Y., SONG, D., XIONG, D., MA, C., HU, X., QU, X., CHEN, L., LI, L., YU, Y., YU, X., ZHANG, Z., ZHAO, C., ZENG, Z., ZHANG, R., YAN, S., WU, T., WU, X., SHU, Y., LEI, J., LI, Y., ZHANG, W., HAYDON, R. C., LUU, H. H., HUANG, A., HE, T.-C. & TANG, H. 2017. Noncanonical Wnt signaling plays an important role in modulating canonical Wnt-regulated stemness, proliferation and terminal differentiation of hepatic progenitors. *Oncotarget,* 8**,** 27105-27119.

FARRINGTON, S. M., TENESA, A., BARNETSON, R., WILTSHIRE, A., PRENDERGAST, J., PORTEOUS, M., CAMPBELL, H. & DUNLOP, M. G. 2005. Germline susceptibility to colorectal cancer due to base-excision repair gene defects. *Am J Hum Genet,* 77**,** 112-9.

FEARNHEAD, N. S., BRITTON, M. P. & BODMER, W. F. 2001. The ABC of APC. *Human Molecular Genetics,* 10**,** 721-733.

FEARON, E. R. & VOGELSTEIN, B. 1990. A genetic model for colorectal tumorigenesis. *Cell,* 61**,** 759-767.

FELTON, K. E. A., GILCHRIST, D. M. & ANDREW, S. E. 2007. Constitutive deficiency in DNA mismatch repair: Is it time for Lynch III? *Clinical Genetics,* 71**,** 499-500.

FISHEL, R., LESCOE, M. K., RAO, M. R. S., COPELAND, N. G., JENKINS, N. A., GARBER, J., KANE, M. & KOLODNER, R. 1993. The human mutator gene homolog MSH2 and its association with hereditary nonpolyposis colon cancer. *Cell,* 75**,** 1027-1038.

FLYNN, A., BENN, D., CLIFTON-BLIGH, R., ROBINSON, B., TRAINER, A. H., JAMES, P., HOGG, A., WALDECK, K., GEORGE, J., LI, J., FOX, S. B., GILL, A. J., MCARTHUR, G., HICKS, R. J. & TOTHILL, R. W. 2015. The genomic landscape of phaeochromocytoma. *The Journal of Pathology,* 236**,** 78-89.

FODDE, R., KUIPERS, J., ROSENBERG, C., SMITS, R., KIELMAN, M., GASPAR, C., VAN ES, J. H., BREUKEL, C., WIEGANT, J., GILES, R. H. & CLEVERS, H. 2001. Mutations in the APC tumour suppressor gene cause chromosomal instability. *Nature Cell Biology,* 3**,** 433-438.

FOKKEMA, I. F. A. C. 2011. *LOVD[3] EP300 gene homepage* [Online]. Available: https://databases.lovd.nl/shared/genes/EP300 [Accessed 23rd November 2017].

FOKKEMA, I. F. A. C., TASCHNER, P. E. M., SCHAAFSMA, G. C. P., CELLI, J., LAROS, J. F. J. & DEN DUNNEN, J. T. 2011. LOVD v.2.0: the next generation in gene variant databases. *Human Mutation,* 32**,** 557-563.

FORBES, S. A., BEARE, D., BOUTSELAKIS, H., BAMFORD, S., BINDAL, N., TATE, J., COLE, C. G., WARD, S., DAWSON, E., PONTING, L., STEFANCSIK, R., HARSHA, B., KOK, C. Y., JIA, M., JUBB, H., SONDKA, Z., THOMPSON, S., DE, T. & CAMPBELL, P. J. 2017. COSMIC: somatic cancer genetics at high-resolution. *Nucleic Acids Research,* 45**,** D777-D783.

FROGGATT, N. J., GREEN, J., BRASSETT, C., EVANS, D. G. R., BISHOP, D. T., KOLODNER, R. & MAHER, E. R. 1999. A common MSH2 mutation in English and North American HNPCC families: origin, phenotypic expression, and sex specific differences in colorectal cancer. *Journal of Medical Genetics,* 36**,** 97-102.

GALA, M. K., MIZUKAMI, Y., LE, L. P., MORIICHI, K., AUSTIN, T., YAMAMOTO, M., LAUWERS, G. Y., BARDEESY, N. & CHUNG, D. C. 2014a. Germline Mutations in Oncogene-Induced Senescence Pathways Are Associated With Multiple Sessile Serrated Adenomas. *Gastroenterology,* 146**,** 520-529.e6.

GALA, M. K., MIZUKAMI, Y., LE, L. P., MORIICHI, K., AUSTIN, T., YAMAMOTO, M., LAUWERS, G. Y., BARDEESY, N. & CHUNG, D. C. 2014b. Germline mutations in oncogene-induced senescence pathways are associated with multiple sessile serrated adenomas. *Gastroenterology,* 146**,** 520-9.

GARA, S. K., JIA, L., MERINO, M. J., AGARWAL, S. K., ZHANG, L., CAM, M., PATEL, D. & KEBEBEW, E. 2015. Germline HABP2 Mutation Causing Familial Nonmedullary Thyroid Cancer. *New England Journal of Medicine,* 373**,** 448-455.

GARCIA-SANZ, P., QUINTANILLA, A., LAFITA, M. C., MORENO-BUENO, G., GARCÍA-GUTIERREZ, L., TABOR, V., VARELA, I., SHIIO, Y., LARSSON, L. G., PORTILLO, F. & LEON, J. 2014. Sin3b Interacts with Myc and decreases Myc levels. *Journal of Biological Chemistry,* 289**,** 22221-22236.

GARRE, P., MARTIN, L., SANZ, J., ROMERO, A., TOSAR, A., BANDO, I., LLOVET, P., DIAQUE, P., GARCIA-PAREDES, B., DIAZ-RUBIO, E., DE LA HOYA, M. & CALDES, T. 2015. BRCA2 gene: a candidate for clinical testing in familial colorectal cancer type X. *Clin Genet,* 87**,** 582-7.

GENOMES PROJECT, C., ABECASIS, G. R., AUTON, A., BROOKS, L. D., DEPRISTO, M. A., DURBIN, R. M., HANDSAKER, R. E., KANG, H. M., MARTH, G. T. & MCVEAN, G. A. 2012. An integrated map of genetic variation from 1,092 human genomes. *Nature,* 491**,** 56-65.

GEOFFROY-PEREZ, B., JANIN, N., OSSIAN, K., LAUGE, A., CROQUETTE, M. F., GRISCELLI, C., DEBRE, M., BRESSAC-DE-PAILLERETS, B., AURIAS, A., STOPPA-LYONNET, D. & ANDRIEU, N. 2001. Cancer risk in heterozygotes for ataxia-telangiectasia. *Int J Cancer,* 93**,** 288-93.

GIARDINE, B., RIEMER, C., HARDISON, R. C., BURHANS, R., ELNITSKI, L., SHAH, P., ZHANG, Y., BLANKENBERG, D., ALBERT, I., TAYLOR, J., MILLER, W., KENT, W. J. & NEKRUTENKO, A. 2005. Galaxy: a platform for interactive large-scale genome analysis. *Genome Res,* 15**,** 1451-5.

GIRÁLDEZ, M. D., BALAGUER, F., BUJANDA, L., CUATRECASAS, M., MUÑOZ, J., ALONSO-ESPINACO, V., LARZABAL, M., PETIT, A., GONZALO, V., OCAÑA, T., MOREIRA, L., ENRÍQUEZ-NAVASCUÉS, J. M., BOLAND, C. R., GOEL, A., CASTELLS, A. & CASTELLVÍ-BEL, S. 2010. MSH6 and MUTYH Deficiency Is a Frequent Event in Early-Onset Colorectal Cancer. *Clinical Cancer Research,* 16**,** 5402-5413.

GOLDBERG, Y., HALPERN, N., HUBERT, A., ADLER, S. N., COHEN, S., PLESSER-DUVDEVANI, M., PAPPO, O., SHAAG, A. & MEINER, V. 2015. Mutated MCM9 is associated with predisposition to hereditary mixed polyposis and colorectal cancer in addition to primary ovarian failure. *Cancer Genet,* 208**,** 621-4.

GOLMARD, L., DELNATTE, C., LAUGÉ, A., MONCOUTIER, V., LEFOL, C., ABIDALLAH, K., TENREIRO, H., COPIGNY, F., GIRAUDEAU, M., GUY, C., BARBAROUX, C., AMORIM, G., BRIAUX, A., GUIBERT, V., TARABEUX, J., CAPUTO, S., COLLET, A., GESTA, P., INGSTER, O., STERN, M. H., ROULEAU, E., DE PAUW, A., GAUTHIER-VILLARS, M., BUECHER, B., BÉZIEAU, S., STOPPA-LYONNET, D. & HOUDAYER, C. 2016. Breast and ovarian cancer predisposition due to de novo BRCA1 and BRCA2 mutations. *Oncogene,* 35**,** 1324-1327.

GONZÁLEZ-PÉREZ, A. & LÓPEZ-BIGAS, N. 2011. Improving the Assessment of the Outcome of Nonsynonymous SNVs with a Consensus Deleteriousness Score, Condel. *American Journal of Human Genetics,* 88**,** 440-449.

GREENMAN, C., STEPHENS, P., SMITH, R., DALGLIESH, G. L., HUNTER, C., BIGNELL, G., DAVIES, H., TEAGUE, J., BUTLER, A., STEVENS, C., EDKINS, S., O'MEARA, S., VASTRIK, I., SCHMIDT, E. E., AVIS, T., BARTHORPE, S., BHAMRA, G., BUCK, G., CHOUDHURY, B., CLEMENTS, J., COLE, J., DICKS, E., FORBES, S., GRAY, K., HALLIDAY, K., HARRISON, R., HILLS, K., HINTON, J., JENKINSON, A., JONES, D., MENZIES, A., MIRONENKO, T., PERRY, J., RAINE, K., RICHARDSON, D., SHEPHERD, R., SMALL, A., TOFTS, C., VARIAN, J., WEBB, T., WEST, S., WIDAA, S., YATES, A., CAHILL, D. P., LOUIS, D. N., GOLDSTRAW, P., NICHOLSON, A. G., BRASSEUR, F., LOOIJENGA, L., WEBER, B. L., CHIEW, Y.-E., DEFAZIO, A., GREAVES, M. F., GREEN, A. R., CAMPBELL, P., BIRNEY, E., EASTON, D. F., CHENEVIX-TRENCH, G., TAN, M.-H., KHOO, S. K., TEH, B. T., YUEN, S. T., LEUNG, S. Y., WOOSTER, R., FUTREAL, P. A. & STRATTON, M. R. 2007. Patterns of somatic mutation in human cancer genomes. *Nature,* 446**,** 153.

GROEN, E., ROOS, A., MUNTINGHE, F., ENTING, R., VRIES, J., KLEIBEUKER, J., WITJES, M. H., LINKS, T. & BEEK, A. 2008. Extra-Intestinal Manifestations of Familial Adenomatous Polyposis. *Annals of Surgical Oncology,* 15**,** 2439-2450.

GROLLEMAN, J. E., DE VOER, R. M., ELSAYED, F. A., NIELSEN, M., WEREN, R. D. A., PALLES, C., LIGTENBERG, M. J. L., VOS, J. R., TEN BROEKE, S. W., DE MIRANDA, N. F. C. C., KUIPER, R. A., KAMPING, E. J., JANSEN, E. A. M., VINK-BÖRGER, M. E., POPP, I., LANG, A., SPIER, I., HÜNEBURG, R., JAMES, P. A., LI, N., STANINOVA, M., LINDSAY, H., COCKBURN, D., SPASIC-BOSKOVIC, O., CLENDENNING, M., SWEET, K., CAPELLÁ, G., SJURSEN, W., HØBERG-VETTI, H., JONGMANS, M. C., NEVELING, K., GEURTS VAN KESSEL, A., MORREAU, H., HES, F. J., SIJMONS, R. H., SCHACKERT, H. K., RUIZ-PONTE, C., DYMERSKA, D., LUBINSKI, J., RIVERA, B., FOULKES, W. D., TOMLINSON, I. P., VALLE, L., BUCHANAN, D. D., KENWRICK, S., ADLARD, J., DIMOVSKI, A. J., CAMPBELL, I. G., ARETZ, S., SCHINDLER, D., VAN WEZEL, T., HOOGERBRUGGE, N. & KUIPER, R. P. 2019.

Mutational Signature Analysis Reveals NTHL1 Deficiency to Cause a Multi-tumor Phenotype. *Cancer Cell,* 35**,** 256-266.e5.

GROVER, S., KASTRINOS, F., STEYERBERG, E. W. & ET AL. 2012. Prevalence and phenotypes of apc and mutyh mutations in patients with multiple colorectal adenomas. *JAMA,* 308**,** 485-492.

GROVES, A., GLEESON, M. & SPIGELMAN, A. D. 2019. NTHL1-associate polyposis: first Australian case report. *Familial Cancer,* 18**,** 179-182.

GRYFE, R., DI NICOLA, N., GEETA, L., GALLINGER, S. & REDSTON, M. 1999. Inherited colorectal polyposis and cancer risk of the APC I1307K polymorphism. *American Journal of Human Genetics,* 64**,** 378-384.

GUARINOS, C., JUAREZ, M., EGOAVIL, C., RODRIGUEZ-SOLER, M., PEREZ-CARBONELL, L., SALAS, R., CUBIELLA, J., RODRIGUEZ-MORANTA, F., DE-CASTRO, L., BUJANDA, L., SERRADESANFERM, A., NICOLAS-PEREZ, D., HERRAIZ, M., FERNANDEZ-BANARES, F., HERREROS-DE-TEJADA, A., AGUIRRE, E., BALMANA, J., RINCON, M. L., PIZARRO, A., POLO-ORTIZ, F., CASTILLEJO, A., ALENDA, C., PAYA, A., SOTO, J. L. & JOVER, R. 2014. Prevalence and characteristics of MUTYH-associated polyposis in patients with multiple adenomatous and serrated polyps. *Clin Cancer Res,* 20**,** 1158-68.

GYLFE, A. E., KATAINEN, R., KONDELIN, J., TANSKANEN, T., CAJUSO, T., HÄNNINEN, U., TAIPALE, J., TAIPALE, M., RENKONEN-SINISALO, L., JÄRVINEN, H., MECKLIN, J.-P., KILPIVAARA, O., PITKÄNEN, E., VAHTERISTO, P., TUUPANEN, S., KARHU, A. & AALTONEN, L. A. 2013a. Eleven Candidate Susceptibility Genes for Common Familial Colorectal Cancer. *PLoS Genetics,* 9**,** e1003876.

GYLFE, A. E., KATAINEN, R., KONDELIN, J., TANSKANEN, T., CAJUSO, T., HANNINEN, U., TAIPALE, J., TAIPALE, M., RENKONEN-SINISALO, L., JARVINEN, H., MECKLIN, J. P., KILPIVAARA, O., PITKANEN, E., VAHTERISTO, P., TUUPANEN, S., KARHU, A. & AALTONEN, L. A. 2013b. Eleven candidate susceptibility genes for common familial colorectal cancer. *PLoS Genet,* 9**,** e1003876.

HABRAKEN, Y., SUNG, P., PRAKASH, L. & PRAKASH, S. 1998. ATP-dependent assembly of a ternary complex consisting of a DNA mismatch and the yeast MSH2-MSH6 and MLH1-PMS1 protein complexes. *J Biol Chem,* 273**,** 9837-41.

HAHN, M.-M., VREEDE, L., BEMELMANS, S. A. S. A., VAN DER LOOIJ, E., VAN KESSEL, A. G., SCHACKERT, H. K., LIGTENBERG, M. J. L., HOOGERBRUGGE, N., KUIPER, R. P. & DE VOER, R. M. 2016. Prevalence of germline mutations in the spindle assembly checkpoint gene BUB1B in individuals with early-onset colorectal cancer. *Genes, Chromosomes and Cancer,* 55**,** 855-863.

HAIMAN, C. A., LE MARCHAND, L., YAMAMATO, J., STRAM, D. O., SHENG, X., KOLONEL, L. N., WU, A. H., REICH, D. & HENDERSON, B. E. 2007. A common genetic risk factor for colorectal and prostate cancer. *Nat Genet,* 39**,** 954-6.

HAMPEL, H., FRANKEL, W. L., MARTIN, E., ARNOLD, M., KHANDUJA, K., KUEBLER, P., NAKAGAWA, H., SOTAMAA, K., PRIOR, T. W., WESTMAN, J., PANESCU, J., FIX, D., LOCKMAN, J., COMERAS, I. & DE LA CHAPELLE, A. 2005. Screening for the Lynch syndrome (hereditary nonpolyposis colorectal cancer). *N Engl J Med,* 352**,** 1851-60.

HANSEN, M. F., JOHANSEN, J., BJORNEVOLL, I., SYLVANDER, A. E., STEINSBEKK, K. S., SAETROM, P., SANDVIK, A. K., DRABLOS, F. & SJURSEN, W. 2015. A novel POLE mutation associated with cancers of colon, pancreas, ovaries and small intestine. *Fam Cancer,* 14**,** 437-48.

HAWKINS, N. J., LEE, J. H. F., WONG, J. J. L., KWOK, C.-T., WARD, R. L. & HITCHINS, M. P. 2009. MGMT methylation is associated primarily with the germline C>T SNP (rs16906252) in colorectal cancer and normal colonic mucosa. *Modern Pathology,* 22**,** 1588.

HEARLE, N., SCHUMACHER, V., MENKO, F. H., OLSCHWANG, S., BOARDMAN, L. A., GILLE, J. J., KELLER, J. J., WESTERMAN, A. M., SCOTT, R. J., LIM, W., TRIMBATH, J. D., GIARDIELLO,

F. M., GRUBER, S. B., OFFERHAUS, G. J., DE ROOIJ, F. W., WILSON, J. H., HANSMANN, A., MOSLEIN, G., ROYER-POKORA, B., VOGEL, T., PHILLIPS, R. K., SPIGELMAN, A. D. & HOULSTON, R. S. 2006. Frequency and spectrum of cancers in the Peutz-Jeghers syndrome. *Clin Cancer Res,* 12**,** 3209-15.

HEISKANEN, I., LUOSTARINEN, T. & JARVINEN, H. J. 2000a. Impact of screening examinations on survival in familial adenomatous polyposis. *Scand J Gastroenterol,* 35**,** 1284-7.

HEISKANEN, I., LUOSTARINEN, T. & JÄRVINEN, H. J. 2000b. Impact of screening examinations on survival in familial adenomatous polyposis. *Scandinavian Journal of Gastroenterology,* 35**,** 1284-1287.

HEMMINKI, A., MARKIE, D., TOMLINSON, I., AVIZIENYTE, E., ROTH, S., LOUKOLA, A., BIGNELL, G., WARREN, W., AMINOFF, M., HOGLUND, P., JARVINEN, H., KRISTO, P., PELIN, K., RIDANPAA, M., SALOVAARA, R., TORO, T., BODMER, W., OLSCHWANG, S., OLSEN, A. S., STRATTON, M. R., DE LA CHAPELLE, A. & AALTONEN, L. A. 1998. A serine/threonine kinase gene defective in Peutz-Jeghers syndrome. *Nature,* 391**,** 184-187.

HERMAN, J. G., UMAR, A., POLYAK, K., GRAFF, J. R., AHUJA, N., ISSA, J.-P. J., MARKOWITZ, S., WILLSON, J. K. V., HAMILTON, S. R., KINZLER, K. W., KANE, M. F., KOLODNER, R. D., VOGELSTEIN, B., KUNKEL, T. A. & BAYLIN, S. B. 1998. Incidence and functional consequences of hMLH1 promoter hypermethylation in colorectal carcinoma. *Proceedings of the National Academy of Sciences,* 95**,** 6870-6875.

HERRERA, L., KAKATI, S., GIBAS, L., PIETRZAK, E., SANDBERG, A. A. & REYNOLDS, J. F. 1986. Gardner syndrome in a man with an interstitial deletion of 5q. *American Journal of Medical Genetics,* 25**,** 473-476.

HICKS, W. M., KIM, M. & HABER, J. E. 2010. Increased Mutagenesis and Unique Mutation Signature Associated with Mitotic Gene Conversion. *Science,* 329**,** 82-85.

HILDEBRANDT , F., BENZING , T. & KATSANIS , N. 2011. Ciliopathies. *New England Journal of Medicine,* 364**,** 1533-1543.

HOISCHEN, A., VAN BON, B. W., GILISSEN, C., ARTS, P., VAN LIER, B., STEEHOUWER, M., DE VRIES, P., DE REUVER, R., WIESKAMP, N., MORTIER, G., DEVRIENDT, K., AMORIM, M. Z., REVENCU, N., KIDD, A., BARBOSA, M., TURNER, A., SMITH, J., OLEY, C., HENDERSON, A., HAYES, I. M., THOMPSON, E. M., BRUNNER, H. G., DE VRIES, B. B. & VELTMAN, J. A. 2010. De novo mutations of SETBP1 cause Schinzel-Giedion syndrome. *Nat Genet,* 42**,** 483-5.

HOLTER, S., BORGIDA, A., DODD, A., GRANT, R., SEMOTIUK, K., HEDLEY, D., DHANI, N., NAROD, S., AKBARI, M., MOORE, M. & GALLINGER, S. 2015. Germline BRCA Mutations in a Large Clinic-Based Cohort of Patients With Pancreatic Adenocarcinoma. *Journal of Clinical Oncology,* 33**,** 3124-3129.

HOULSTON, R. S., CHEADLE, J., DOBBINS, S. E., TENESA, A., JONES, A. M., HOWARTH, K., SPAIN, S. L., BRODERICK, P., DOMINGO, E., FARRINGTON, S., PRENDERGAST, J. G., PITTMAN, A. M., THEODORATOU, E., SMITH, C. G., OLVER, B., WALTHER, A., BARNETSON, R. A., CHURCHMAN, M., JAEGER, E. E., PENEGAR, S., BARCLAY, E., MARTIN, L., GORMAN, M., MAGER, R., JOHNSTONE, E., MIDGLEY, R., NIITTYMAKI, I., TUUPANEN, S., COLLEY, J., IDZIASZCZYK, S., THOMAS, H. J., LUCASSEN, A. M., EVANS, D. G., MAHER, E. R., MAUGHAN, T., DIMAS, A., DERMITZAKIS, E., CAZIER, J. B., AALTONEN, L. A., PHAROAH, P., KERR, D. J., CARVAJAL-CARMONA, L. G., CAMPBELL, H., DUNLOP, M. G. & TOMLINSON, I. P. 2010. Meta-analysis of three genome-wide association studies identifies susceptibility loci for colorectal cancer at 1q41, 3q26.2, 12q13.13 and 20q13.33. *Nat Genet,* 42**,** 973-7.

HOULSTON, R. S., WEBB, E., BRODERICK, P., PITTMAN, A. M., DI BERNARDO, M. C., LUBBE, S., CHANDLER, I., VIJAYAKRISHNAN, J., SULLIVAN, K., PENEGAR, S., CARVAJAL-CARMONA, L., HOWARTH, K., JAEGER, E., SPAIN, S. L., WALTHER, A., BARCLAY, E., MARTIN, L., GORMAN, M., DOMINGO, E., TEIXEIRA, A. S., KERR, D., CAZIER, J. B., NIITTYMAKI, I.,

TUUPANEN, S., KARHU, A., AALTONEN, L. A., TOMLINSON, I. P., FARRINGTON, S. M., TENESA, A., PRENDERGAST, J. G., BARNETSON, R. A., CETNARSKYJ, R., PORTEOUS, M. E., PHAROAH, P. D., KOESSLER, T., HAMPE, J., BUCH, S., SCHAFMAYER, C., TEPEL, J., SCHREIBER, S., VOLZKE, H., CHANG-CLAUDE, J., HOFFMEISTER, M., BRENNER, H., ZANKE, B. W., MONTPETIT, A., HUDSON, T. J., GALLINGER, S., CAMPBELL, H. & DUNLOP, M. G. 2008. Meta-analysis of genome-wide association data identifies four new susceptibility loci for colorectal cancer. *Nat Genet,* 40**,** 1426-35.

HOWE, J. R., BAIR, J. L., SAYED, M. G., ANDERSON, M. E., MITROS, F. A., PETERSEN, G. M., VELCULESCU, V. E., TRAVERSO, G. & VOGELSTEIN, B. 2001a. Germline mutations of the gene encoding bone morphogenetic protein receptor 1A in juvenile polyposis. *Nat Genet,* 28**,** 184-7.

HOWE, J. R., BAIR, J. L., SAYED, M. G., ANDERSON, M. E., MITROS, F. A., PETERSEN, G. M., VELCULESCU, V. E., TRAVERSO, G. & VOGELSTEIN, B. 2001b. Germline mutations of the gene encoding bone morphogenetic protein receptor 1A in juvenile polyposis. *Nat Genet,* 28**,** 184-187.

HOWE, J. R., ROTH, S., RINGOLD, J. C., SUMMERS, R. W., JÄRVINEN, H. J., SISTONEN, P., TOMLINSON, I. P. M., HOULSTON, R. S., BEVAN, S., MITROS, F. A., STONE, E. M. & AALTONEN, L. A. 1998. Mutations in the SMAD4/DPC4 Gene in Juvenile Polyposis. *Science,* 280**,** 1086-1088.

HOWE, J. R., SAYED, M. G., AHMED, A. F., RINGOLD, J., LARSEN-HAIDLE, J., MERG, A., MITROS, F. A., VACCARO, C. A., PETERSEN, G. M., GIARDIELLO, F. M., TINLEY, S. T., AALTONEN, L. A. & LYNCH, H. T. 2004. The prevalence of *MADH4* and *BMPR1A* mutations in juvenile polyposis and absence of *BMPR2, BMPR1B,* and *ACVR1* mutations. *Journal of Medical Genetics,* 41**,** 484-491.

HUTTER, C. M., SLATTERY, M. L., DUGGAN, D. J., MUEHLING, J., CURTIN, K., HSU, L., BERESFORD, S. A., RAJKOVIC, A., SARTO, G. E., MARSHALL, J. R., HAMMAD, N., WALLACE, R., MAKAR, K. W., PRENTICE, R. L., CAAN, B. J., POTTER, J. D. & PETERS, U. 2010. Characterization of the association between 8q24 and colon cancer: gene-environment exploration and meta-analysis. *BMC Cancer,* 10**,** 670.

JAEGER, E., LEEDHAM, S., LEWIS, A., SEGDITSAS, S., BECKER, M., CUADRADO, P. R., DAVIS, H., KAUR, K., HEINIMANN, K., HOWARTH, K., EAST, J., TAYLOR, J., THOMAS, H. & TOMLINSON, I. 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**,** 699-703.

JAEGER, E., WEBB, E., HOWARTH, K., CARVAJAL-CARMONA, L., ROWAN, A., BRODERICK, P., WALTHER, A., SPAIN, S., PITTMAN, A., KEMP, Z., SULLIVAN, K., HEINIMANN, K., LUBBE, S., DOMINGO, E., BARCLAY, E., MARTIN, L., GORMAN, M., CHANDLER, I., VIJAYAKRISHNAN, J., WOOD, W., PAPAEMMANUIL, E., PENEGAR, S., QURESHI, M., FARRINGTON, S., TENESA, A., CAZIER, J. B., KERR, D., GRAY, R., PETO, J., DUNLOP, M., CAMPBELL, H., THOMAS, H., HOULSTON, R. & TOMLINSON, I. 2008. Common genetic variants at the CRAC1 (HMPS) locus on chromosome 15q13.3 influence colorectal cancer risk. *Nat Genet,* 40**,** 26-8.

JÄRVINEN*, H. J., AARNIO*, M., MUSTONEN*, H., AKTAN–COLLAN‡, K., AALTONEN‡, L. A., PELTOMÄKI§, P., CHAPELLE§, A. D. L. & MECKLIN‖, J. P. 2000. Controlled 15-year trial on screening for colorectal cancer in families with hereditary nonpolyposis colorectal cancer. *Gastroenterology,* 118**,** 829-834.

JASPERSON, K. W., SAMOWITZ, W. S. & BURT, R. W. 2011. Constitutional mismatch repair-deficiency syndrome presenting as colonic adenomatous polyposis: clues from the skin. *Clin Genet,* 80**,** 394-7.

JASPERSON, K. W., TUOHY, T. M., NEKLASON, D. W. & BURT, R. W. 2010. Hereditary and familial colon cancer. *Gastroenterology,* 138**,** 2044-58.

JASS, J. R. 2007a. Classification of colorectal cancer based on correlation of clinical, morphological and molecular features. *Histopathology,* 50**,** 113-130.

JASS, J. R. 2007b. Gastrointestinal polyposes: clinical, pathological and molecular features. *Gastroenterol Clin North Am,* 36**,** 927-46, viii.

JEMAL, A., BRAY, F., CENTER, M. M., FERLAY, J., WARD, E. & FORMAN, D. 2011. Global cancer statistics. *CA Cancer J Clin,* 61**,** 69-90.

JEMAL, A., SIEGEL, R., WARD, E., HAO, Y., XU, J. & THUN, M. J. 2009. Cancer Statistics, 2009. *CA: A Cancer Journal for Clinicians,* 59**,** 225-249.

JENNE, D. E., REIMANN, H., NEZU, J. I., FRIEDEL, W., LOFF, S., JESCHKE, R., MÜLLER, O., BACK, W. & ZIMMER, M. 1998. Peutz-Jeghers syndrome is caused by mutations in a novel serine threonine kinase. *Nature Genetics,* 18**,** 38-43.

JIA, W. H., ZHANG, B., MATSUO, K., SHIN, A., XIANG, Y. B., JEE, S. H., KIM, D. H., REN, Z., CAI, Q., LONG, J., SHI, J., WEN, W., YANG, G., DELAHANTY, R. J., JI, B. T., PAN, Z. Z., MATSUDA, F., GAO, Y. T., OH, J. H., AHN, Y. O., PARK, E. J., LI, H. L., PARK, J. W., JO, J., JEONG, J. Y., HOSONO, S., CASEY, G., PETERS, U., SHU, X. O., ZENG, Y. X. & ZHENG, W. 2013. Genome-wide association analyses in East Asians identify new susceptibility loci for colorectal cancer. *Nat Genet,* 45**,** 191-6.

JOHNS, L. E. & HOULSTON, R. S. 2001. A systematic review and meta-analysis of familial colorectal cancer risk. *Am J Gastroenterol,* 96**,** 2992-3003.

JOHNSON, V. L., SCOTT, M. I. F., HOLT, S. V., HUSSEIN, D. & TAYLOR, S. S. 2004. Bub1 is required for kinetochore localization of BubR1, Cenp-E, Cenp-F and Mad2, and chromosome congression. *Journal of Cell Science,* 117**,** 1577-1589.

JONES, N., VOGT, S., NIELSEN, M., CHRISTIAN, D., WARK, P. A., ECCLES, D., EDWARDS, E., EVANS, D. G., MAHER, E. R., VASEN, H. F., HES, F. J., ARETZ, S. & SAMPSON, J. R. 2009a. Increased Colorectal Cancer Incidence in Obligate Carriers of Heterozygous Mutations in MUTYH. *Gastroenterology,* 137**,** 489-494.e1.

JONES, S., EMMERSON, P., MAYNARD, J., BEST, J. M., JORDAN, S., WILLIAMS, G. T., SAMPSON, J. R. & CHEADLE, J. P. 2002. Biallelic germline mutations in MYH predispose to multiple colorectal adenoma and somatic G:C-->T:A mutations. *Hum Mol Genet,* 11**,** 2961-7.

JONES, S., HRUBAN, R. H., KAMIYAMA, M., BORGES, M., ZHANG, X., PARSONS, D. W., LIN, J. C.-H., PALMISANO, E., BRUNE, K., JAFFEE, E. M., IACOBUZIO-DONAHUE, C. A., MAITRA, A., PARMIGIANI, G., KERN, S. E., VELCULESCU, V. E., KINZLER, K. W., VOGELSTEIN, B., ESHLEMAN, J. R., GOGGINS, M. & KLEIN, A. P. 2009b. Exomic Sequencing Identifies PALB2 as a Pancreatic Cancer Susceptibility Gene. *Science,* 324**,** 217-217.

JONES, S., HRUBAN, R. H., KAMIYAMA, M., BORGES, M., ZHANG, X., PARSONS, D. W., LIN, J. C.-H., PALMISANO, E., BRUNE, K., JAFFEE, E. M., IACOBUZIO-DONAHUE, C. A., MAITRA, A., PARMIGIANI, G., KERN, S. E., VELCULESCU, V. E., KINZLER, K. W., VOGELSTEIN, B., ESHLEMAN, J. R., GOGGINS, M. & KLEIN, A. P. 2009c. Exomic Sequencing Identifies PALB2 as a Pancreatic Cancer Susceptibility Gene. *Science,* 324**,** 217-217.

JOSLYN, G., CARLSON, M., THLIVERIS, A., ALBERTSEN, H., GELBERT, L., SAMOWITZ, W., GRODEN, J., STEVENS, J., SPIRIO, L., ROBERTSON, M. & ET AL. 1991. Identification of deletion mutations and three new genes at the familial polyposis locus. *Cell,* 66**,** 601-13.

KADYROV, F. A., DZANTIEV, L., CONSTANTIN, N. & MODRICH, P. 2006. Endonucleolytic function of MutLalpha in human mismatch repair. *Cell,* 126**,** 297-308.

KADYROV, F. A., HOLMES, S. F., ARANA, M. E., LUKIANOVA, O. A., O'DONNELL, M., KUNKEL, T. A. & MODRICH, P. 2007. Saccharomyces cerevisiae MutLalpha is a mismatch repair endonuclease. *J Biol Chem,* 282**,** 37181-90.

KALADY, M. F., JARRAR, A., LEACH, B., LAGUARDIA, L., O'MALLEY, M., ENG, C. & CHURCH, J. M. 2011. Defining phenotypes and cancer risk in hyperplastic polyposis syndrome. *Dis Colon Rectum,* 54**,** 164-70.

KATOH, M. & KATOH, M. 2007. Notch signaling in gastrointestinal tract (Review). *International Journal of Oncology,* 30**,** 247-251.

KEMPERS, M. J., KUIPER, R. P., OCKELOEN, C. W., CHAPPUIS, P. O., HUTTER, P., RAHNER, N., SCHACKERT, H. K., STEINKE, V., HOLINSKI-FEDER, E., MORAK, M., KLOOR, M., BUTTNER, R., VERWIEL, E. T., VAN KRIEKEN, J. H., NAGTEGAAL, I. D., GOOSSENS, M., VAN DER POST, R. S., NIESSEN, R. C., SIJMONS, R. H., KLUIJT, I., HOGERVORST, F. B., LETER, E. M., GILLE, J. J., AALFS, C. M., REDEKER, E. J., HES, F. J., TOPS, C. M., VAN NESSELROOIJ, B. P., VAN GIJN, M. E., GOMEZ GARCIA, E. B., ECCLES, D. M., BUNYAN, D. J., SYNGAL, S., STOFFEL, E. M., CULVER, J. O., PALOMARES, M. R., GRAHAM, T., VELSHER, L., PAPP, J., OLAH, E., CHAN, T. L., LEUNG, S. Y., VAN KESSEL, A. G., KIEMENEY, L. A., HOOGERBRUGGE, N. & LIGTENBERG, M. J. 2011. Risk of colorectal and endometrial cancers in EPCAM deletion-positive Lynch syndrome: a cohort study. *Lancet Oncol,* 12**,** 49-55.

KIM, D. W., KIM, I. J., KANG, H. C., JANG, S. G., KIM, K., YOON, H. J., AHN, S. A., HAN, S. Y., HONG, S. H., HWANG, J. A., SOHN, D. K., JEONG, S. Y., CHOI, H. S., HONG, C. W., LIM, S. B. & PARK, J. G. 2007. Germline mutations of the MYH gene in Korean patients with multiple colorectal adenomas. *Int J Colorectal Dis,* 22**,** 1173-8.

KIRCHER, M., WITTEN, D. M., JAIN, P., O'ROAK, B. J., COOPER, G. M. & SHENDURE, J. 2014. A general framework for estimating the relative pathogenicity of human genetic variants. *Nat Genet,* 46**,** 310-5.

KISSELEVA, M. V., WILSON, M. P. & MAJERUS, P. W. 2000. The isolation and characterization of a cDNA encoding phospholipid- specific inositol polyphosphate 5-phosphatase. *Journal of Biological Chemistry,* 275**,** 20110-20116.

KNUDSEN, A. L., BULOW, S., TOMLINSON, I., MOSLEIN, G., HEINIMANN, K., CHRISTENSEN, I. J. & GROUP, A. S. 2010. Attenuated familial adenomatous polyposis: results from an international collaborative study. *Colorectal Dis,* 12**,** e243-9.

KRYUKOV, G. V., PENNACCHIO, L. A. & SUNYAEV, S. R. 2007. Most Rare Missense Alleles Are Deleterious in Humans: Implications for Complex Disease and Association Studies. *The American Journal of Human Genetics,* 80**,** 727-739.

KU, C.-S., COOPER, D. N., WU, M., ROUKOS, D. H., PAWITAN, Y., SOONG, R. & IACOPETTA, B. 2012. Gene discovery in familial cancer syndromes by exome sequencing: prospects for the elucidation of familial colorectal cancer type X. *Mod Pathol,* 25**,** 1055-1068.

KU, C. S., NAIDOO, N. & PAWITAN, Y. 2011. Revisiting Mendelian disorders through exome sequencing. *Hum Genet,* 129**,** 351-70.

KUNKEL, T. A. & ERIE, D. A. 2005. DNA mismatch repair. *Annu Rev Biochem,* 74**,** 681-710.

LAKEN, S. J., PETERSEN, G. M., GRUBER, S. B., ODDOUX, C., OSTRER, H., GIARDIELLO, F. M., HAMILTON, S. R., HAMPEL, H., MARKOWITZ, A., KLIMSTRA, D., JHANWAR, S., WINAWER, S., OFFIT, K., LUCE, M. C., KINZLER, K. W. & VOGELSTEIN, B. 1997. Familial colorectal cancer in Ashkenazim due to a hypermutable tract in APC. *Nat Genet,* 17**,** 79-83.

LANDRUM, M. J., LEE, J. M., BENSON, M., BROWN, G., CHAO, C., CHITIPIRALLA, S., GU, B., HART, J., HOFFMAN, D., HOOVER, J., JANG, W., KATZ, K., OVETSKY, M., RILEY, G., SETHI, A., TULLY, R., VILLAMARIN-SALOMON, R., RUBINSTEIN, W. & MAGLOTT, D. R. 2016. ClinVar: public archive of interpretations of clinically relevant variants. *Nucleic Acids Res,* 44**,** D862-8.

LANDRUM, M. J., LEE, J. M., BENSON, M., BROWN, G. R., CHAO, C., CHITIPIRALLA, S., GU, B., HART, J., HOFFMAN, D., JANG, W., KARAPETYAN, K., KATZ, K., LIU, C., MADDIPATLA, Z., MALHEIRO, A., MCDANIEL, K., OVETSKY, M., RILEY, G., ZHOU, G., HOLMES, J. B., KATTMAN, B. L. & MAGLOTT, D. R. 2018. ClinVar: improving access to variant interpretations and supporting evidence. *Nucleic Acids Res,* 46**,** D1062-D1067.

LANSPA, S. J., AHNEN, D. J. & LYNCH, H. T. 2012. Editorial: Serrated Polyposis: The Last (or Only the Latest[quest]) Frontier of Familial Polyposis[quest]. *Am J Gastroenterol,* 107**,** 779-781.

LARSON, A. M., SS, H., V, D., AO, S. R., MG, H., CR, F., U, S. & EA, T. 2012. Pancreatic neuroendocrine tumors in patients with tuberous sclerosis complex. *Clinical Genetics,* 82**,** 558-563.

LE, D. T., URAM, J. N., WANG, H., BARTLETT, B. R., KEMBERLING, H., EYRING, A. D., SKORA, A. D., LUBER, B. S., AZAD, N. S., LAHERU, D., BIEDRZYCKI, B., DONEHOWER, R. C., ZAHEER, A., FISHER, G. A., CROCENZI, T. S., LEE, J. J., DUFFY, S. M., GOLDBERG, R. M., DE LA CHAPELLE, A., KOSHIJI, M., BHAIJEE, F., HUEBNER, T., HRUBAN, R. H., WOOD, L. D., CUKA, N., PARDOLL, D. M., PAPADOPOULOS, N., KINZLER, K. W., ZHOU, S., CORNISH, T. C., TAUBE, J. M., ANDERS, R. A., ESHLEMAN, J. R., VOGELSTEIN, B. & DIAZ, L. A. J. 2015. PD-1 Blockade in Tumors with Mismatch-Repair Deficiency. *New England Journal of Medicine,* 372**,** 2509-2520.

LEK, M., KARCZEWSKI, K. J., MINIKEL, E. V., SAMOCHA, K. E., BANKS, E., FENNELL, T., O'DONNELL-LURIA, A. H., WARE, J. S., HILL, A. J., CUMMINGS, B. B., TUKIAINEN, T., BIRNBAUM, D. P., KOSMICKI, J. A., DUNCAN, L. E., ESTRADA, K., ZHAO, F., ZOU, J., PIERCE-HOFFMAN, E., BERGHOUT, J., COOPER, D. N., DEFLAUX, N., DEPRISTO, M., DO, R., FLANNICK, J., FROMER, M., GAUTHIER, L., GOLDSTEIN, J., GUPTA, N., HOWRIGAN, D., KIEZUN, A., KURKI, M. I., MOONSHINE, A. L., NATARAJAN, P., OROZCO, L., PELOSO, G. M., POPLIN, R., RIVAS, M. A., RUANO-RUBIO, V., ROSE, S. A., RUDERFER, D. M., SHAKIR, K., STENSON, P. D., STEVENS, C., THOMAS, B. P., TIAO, G., TUSIE-LUNA, M. T., WEISBURD, B., WON, H.-H., YU, D., ALTSHULER, D. M., ARDISSINO, D., BOEHNKE, M., DANESH, J., DONNELLY, S., ELOSUA, R., FLOREZ, J. C., GABRIEL, S. B., GETZ, G., GLATT, S. J., HULTMAN, C. M., KATHIRESAN, S., LAAKSO, M., MCCARROLL, S., MCCARTHY, M. I., MCGOVERN, D., MCPHERSON, R., NEALE, B. M., PALOTIE, A., PURCELL, S. M., SALEHEEN, D., SCHARF, J. M., SKLAR, P., SULLIVAN, P. F., TUOMILEHTO, J., TSUANG, M. T., WATKINS, H. C., WILSON, J. G., DALY, M. J., MACARTHUR, D. G. & EXOME AGGREGATION, C. 2016a. Analysis of protein-coding genetic variation in 60,706 humans. *Nature,* 536**,** 285-291.

LEK, M., KARCZEWSKI, K. J., MINIKEL, E. V., SAMOCHA, K. E., BANKS, E., FENNELL, T., O'DONNELL-LURIA, A. H., WARE, J. S., HILL, A. J., CUMMINGS, B. B., TUKIAINEN, T., BIRNBAUM, D. P., KOSMICKI, J. A., DUNCAN, L. E., ESTRADA, K., ZHAO, F., ZOU, J., PIERCE-HOFFMAN, E., BERGHOUT, J., COOPER, D. N., DEFLAUX, N., DEPRISTO, M., DO, R., FLANNICK, J., FROMER, M., GAUTHIER, L., GOLDSTEIN, J., GUPTA, N., HOWRIGAN, D., KIEZUN, A., KURKI, M. I., MOONSHINE, A. L., NATARAJAN, P., OROZCO, L., PELOSO, G. M., POPLIN, R., RIVAS, M. A., RUANO-RUBIO, V., ROSE, S. A., RUDERFER, D. M., SHAKIR, K., STENSON, P. D., STEVENS, C., THOMAS, B. P., TIAO, G., TUSIE-LUNA, M. T., WEISBURD, B., WON, H. H., YU, D., ALTSHULER, D. M., ARDISSINO, D., BOEHNKE, M., DANESH, J., DONNELLY, S., ELOSUA, R., FLOREZ, J. C., GABRIEL, S. B., GETZ, G., GLATT, S. J., HULTMAN, C. M., KATHIRESAN, S., LAAKSO, M., MCCARROLL, S., MCCARTHY, M. I., MCGOVERN, D., MCPHERSON, R., NEALE, B. M., PALOTIE, A., PURCELL, S. M., SALEHEEN, D., SCHARF, J. M., SKLAR, P., SULLIVAN, P. F., TUOMILEHTO, J., TSUANG, M. T., WATKINS, H. C., WILSON, J. G., DALY, M. J., MACARTHUR, D. G. & EXOME AGGREGATION, C. 2016b. Analysis of protein-coding genetic variation in 60,706 humans. *Nature,* 536**,** 285-91.

LEK, M., KARCZEWSKI, K. J., MINIKEL, E. V., SAMOCHA, K. E., BANKS, E., FENNELL, T., O'DONNELL-LURIA, A. H., WARE, J. S., HILL, A. J., CUMMINGS, B. B., TUKIAINEN, T., BIRNBAUM, D. P., KOSMICKI, J. A., DUNCAN, L. E., ESTRADA, K., ZHAO, F., ZOU, J., PIERCE-HOFFMAN, E., BERGHOUT, J., COOPER, D. N., DEFLAUX, N., DEPRISTO, M., DO, R., FLANNICK, J., FROMER, M., GAUTHIER, L., GOLDSTEIN, J., GUPTA, N., HOWRIGAN,

D., KIEZUN, A., KURKI, M. I., MOONSHINE, A. L., NATARAJAN, P., OROZCO, L., PELOSO, G. M., POPLIN, R., RIVAS, M. A., RUANO-RUBIO, V., ROSE, S. A., RUDERFER, D. M., SHAKIR, K., STENSON, P. D., STEVENS, C., THOMAS, B. P., TIAO, G., TUSIE-LUNA, M. T., WEISBURD, B., WON, H.-H., YU, D., ALTSHULER, D. M., ARDISSINO, D., BOEHNKE, M., DANESH, J., DONNELLY, S., ELOSUA, R., FLOREZ, J. C., GABRIEL, S. B., GETZ, G., GLATT, S. J., HULTMAN, C. M., KATHIRESAN, S., LAAKSO, M., MCCARROLL, S., MCCARTHY, M. I., MCGOVERN, D., MCPHERSON, R., NEALE, B. M., PALOTIE, A., PURCELL, S. M., SALEHEEN, D., SCHARF, J. M., SKLAR, P., SULLIVAN, P. F., TUOMILEHTO, J., TSUANG, M. T., WATKINS, H. C., WILSON, J. G., DALY, M. J., MACARTHUR, D. G. & EXOME AGGREGATION, C. 2016c. Analysis of protein-coding genetic variation in 60,706 humans. *Nature,* 536**,** 285-291.

LEPPERT, M., DOBBS, M., SCAMBLER, P., O'CONNELL, P., NAKAMURA, Y., STAUFFER, D., WOODWARD, S., BURT, R., HUGHES, J., GARDNER, E. & ET, A. 1987. The gene for familial polyposis coli maps to the long arm of chromosome 5. *Science,* 238**,** 1411-1413.

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

LI, N., THOMPSON, E. R., ROWLEY, S. M., MCINERNY, S., DEVEREUX, L., GOODE, D., INVESTIGATORS, L., WONG-BROWN, M. W., SCOTT, R. J., TRAINER, A. H., GORRINGE, K. L., JAMES, P. A. & CAMPBELL, I. G. 2016. Reevaluation of RINT1 as a breast cancer predisposition gene. *Breast Cancer Research and Treatment,* 159**,** 385-392.

LI, W. D., LI, Q. R., XU, S. N., WEI, F. J., YE, Z. J., CHENG, J. K. & CHEN, J. P. 2013. Exome sequencing identifies an MLL3 gene germ line mutation in a pedigree of colorectal cancer and acute myeloid leukemia. *Blood,* 121**,** 1478-9.

LIANG, J., LIN, C., HU, F., WANG, F., ZHU, L., YAO, X., WANG, Y. & ZHAO, Y. 2013. APC polymorphisms and the risk of colorectal neoplasia: a HuGE review and meta-analysis. *Am J Epidemiol,* 177**,** 1169-79.

LIGTENBERG, M. J., KUIPER, R. P., CHAN, T. L., GOOSSENS, M., HEBEDA, K. M., VOORENDT, M., LEE, T. Y., BODMER, D., HOENSELAAR, E., HENDRIKS-CORNELISSEN, S. J., TSUI, W. Y., KONG, C. K., BRUNNER, H. G., VAN KESSEL, A. G., YUEN, S. T., VAN KRIEKEN, J. H., LEUNG, S. Y. & HOOGERBRUGGE, N. 2009. Heritable somatic methylation and inactivation of MSH2 in families with Lynch syndrome due to deletion of the 3' exons of TACSTD1. *Nat Genet,* 41**,** 112-7.

LINDBLOM, A., TANNERGARD, P., WERELIUS, B. & NORDENSKJOLD, M. 1993. Genetic mapping of a second locus predisposing to hereditary non-polyposis colon cancer. *Nat Genet,* 5**,** 279-282.

LINDOR, N. M. 2009. Familial colorectal cancer type X: the other half of hereditary nonpolyposis colon cancer syndrome. *Surg Oncol Clin N Am,* 18**,** 637-45.

LINDOR, N. M., RABE, K., PETERSEN, G. M., HAILE, R., CASEY, G., BARON, J., GALLINGER, S., BAPAT, B., ARONSON, M., HOPPER, J., JASS, J., LEMARCHAND, L., GROVE, J., POTTER, J., NEWCOMB, P., TERDIMAN, J. P., CONRAD, P., MOSLEIN, G., GOLDBERG, R., ZIOGAS, A., ANTON-CULVER, H., DE ANDRADE, M., SIEGMUND, K., THIBODEAU, S. N., BOARDMAN, L. A. & SEMINARA, D. 2005. Lower cancer incidence in Amsterdam-I criteria families without mismatch repair deficiency: Familial colorectal cancer type X. *Journal of the American Medical Association,* 293**,** 1979-1985.

LUBBE, S. J., BERNARDO, M. C. D., CHANDLER, I. P. & HOULSTON, R. S. 2009. Clinical Implications of the Colorectal Cancer Risk Associated With MUTYH Mutation. *Journal of Clinical Oncology,* 27**,** 3975-3980.

LYNCH, H. T. & DE LA CHAPELLE, A. 2003. Hereditary colorectal cancer. *N Engl J Med,* 348**,** 919-32.

LYNCH, H. T., LANSPA, S., SMYRK, T., BOMAN, B., WATSON, P. & LYNCH, J. 1991a. Hereditary nonpolyposis colorectal cancer (Lynch syndromes I & II). Genetics, pathology, natural history, and cancer control, Part I. *Cancer Genet Cytogenet,* 53**,** 143-60.

LYNCH, H. T., LYNCH, P. M., LANSPA, S. J., SNYDER, C. L., LYNCH, J. F. & BOLAND, C. R. 2009. Review of the Lynch syndrome: history, molecular genetics, screening, differential diagnosis, and medicolegal ramifications. *Clin Genet,* 76**,** 1-18.

LYNCH, H. T., SHAW, M. W., MAGNUSON, C. W., LARSEN, A. L. & KRUSH, A. J. 1966. Hereditary factors in cancer. Study of two large midwestern kindreds. *Archives of Internal Medicine,* 117**,** 206-212.

LYNCH, H. T., SMYRK, T., WATSON, P., LANSPA, S. J., BOMAN, B. M., LYNCH, P. M., LYNCH, J. F. & CAVALIERI, J. 1991b. Hereditary colorectal cancer. *Semin Oncol,* 18**,** 337-66.

MACRAE, C. J., MCCULLOCH, R. D., YLANKO, J., DUROCHER, D. & KOCH, C. A. 2008. APLF (C2orf13) facilitates nonhomologous end-joining and undergoes ATM-dependent hyperphosphorylation following ionizing radiation. *DNA Repair,* 7**,** 292-302.

MAJEWSKI, J., SCHWARTZENTRUBER, J., LALONDE, E., MONTPETIT, A. & JABADO, N. 2011. What can exome sequencing do for you? *Journal of Medical Genetics,* 48**,** 580-589.

MARCH, H. N., RUST, A. G., WRIGHT, N. A., TEN HOEVE, J., DE RIDDER, J., ELDRIDGE, M., VAN DER WEYDEN, L., BERNS, A., GADIOT, J., UREN, A., KEMP, R., ARENDS, M. J., WESSELS, L. F. A., WINTON, D. J. & ADAMS, D. J. 2011. Insertional mutagenesis identifies multiple networks of cooperating genes driving intestinal tumorigenesis. *Nature Genetics,* 43**,** 1202.

MARTIN, A. & SCHARFF, M. D. 2002. AID and mismatch repair in antibody diversification. *Nature Reviews. Immunology,* 2**,** 605-614.

MARTIN, E. S., TONON, G., SINHA, R., XIAO, Y., FENG, B., KIMMELMAN, A. C., PROTOPOPOV, A., IVANOVA, E., BRENNAN, C., MONTGOMERY, K., KUCHERLAPATI, R., BAILEY, G., REDSTON, M., CHIN, L. & DEPINHO, R. A. 2007. Common and distinct genomic events in sporadic colorectal cancer and diverse cancer types. *Cancer Res,* 67**,** 10736-43.

MARTIN, M. 2011. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet.journal; Vol 17, No 1: Next Generation Sequencing Data Analysis*.

MARTOMO, S. A., YANG, W. W. & GEARHART, P. J. 2004. A Role for Msh6 But Not Msh3 in Somatic Hypermutation and Class Switch Recombination. *The Journal of Experimental Medicine,* 200**,** 61-68.

MAZOYER, S., DUNNING, A. M., SEROVA, O., DEARDEN, J., PUGET, N., HEALEY, C. S., GAYTHER, S. A., MANGION, J., STRATTON, M. R., LYNCH, H. T., GOLDGAR, D. E., PONDER, B. A. & LENOIR, G. M. 1996. A polymorphic stop codon in BRCA2. *Nat Genet,* 14**,** 253-4.

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

MCLAREN, W., PRITCHARD, B., RIOS, D., CHEN, Y., FLICEK, P. & CUNNINGHAM, F. 2010. Deriving the consequences of genomic variants with the Ensembl API and SNP Effect Predictor. *Bioinformatics,* 26**,** 2069-70.

MECKLIN, J.-P., SIPPONEN, P. & JÄRVINEN, H. J. 1986. Histopathology of colorectal carcinomas and adenomas in cancer family syndrome. *Diseases of the Colon & Rectum,* 29**,** 849-853.

MEHROTRA, P. V., AHEL, D., RYAN, D. P., WESTON, R., WIECHENS, N., KRAEHENBUEHL, R., OWEN-HUGHES, T. & AHEL, I. 2011. DNA Repair Factor APLF Is a Histone Chaperone. *Molecular Cell,* 41**,** 46-55.

MENG, F., ZHANG, L., REN, Y. & MA, Q. 2019. The genomic alterations of lung adenocarcinoma and lung squamous cell carcinoma can explain the differences of their overall survival rates. *Journal of Cellular Physiology,* 234**,** 10918-10925.

MERALDI, P. & SORGER, P. K. 2005. A dual role for Bub1 in the spindle checkpoint and chromosome congression. *The EMBO Journal,* 24**,** 1621-1633.
MERTES, F., ELSHARAWY, A., SAUER, S., VAN HELVOORT, J. M., VAN DER ZAAG, P. J., FRANKE, A., NILSSON, M., LEHRACH, H. & BROOKES, A. J. 2011. Targeted enrichment of genomic DNA regions for next-generation sequencing. *Brief Funct Genomics,* 10**,** 374-86.
MICHAELS, M. L. & MILLER, J. H. 1992. The GO system protects organisms from the mutagenic effect of the spontaneous lesion 8-hydroxyguanine (7,8-dihydro-8-oxoguanine). *J Bacteriol,* 174**,** 6321-5.
MILANI, D., BONARRIGO, F. A., MENNI, F., SPACCINI, L., GERVASINI, C. & ESPOSITO, S. 2016. Hepatoblastoma in Rubinstein-Taybi Syndrome: A Case Report. *Pediatric Blood and Cancer,* 63**,** 572-573.
MIYAKI, M., KONISHI, M., TANAKA, K., KIKUCHI-YANOSHITA, R., MURAOKA, M., YASUNO, M., IGARI, T., KOIKE, M., CHIBA, M. & MORI, T. 1997. Germline mutation of MSH6 as the cause of hereditary nonpolyposis colorectal cancer. *Nat Genet,* 17**,** 271-2.
MODRICH, P. 1991. Mechanisms and biological effects of mismatch repair. *Annu Rev Genet,* 25**,** 229-53.
MØLLER, P., SEPPÄLÄ, T. T., BERNSTEIN, I., HOLINSKI-FEDER, E., SALA, P., GARETH EVANS, D., LINDBLOM, A., MACRAE, F., BLANCO, I., SIJMONS, R. H., JEFFRIES, J., VASEN, H. F. A., BURN, J., NAKKEN, S., HOVIG, E., RØDLAND, E. A., THARMARATNAM, K., DE VOS TOT NEDERVEEN CAPPEL, W. H., HILL, J., WIJNEN, J. T., JENKINS, M. A., GREEN, K., LALLOO, F., SUNDE, L., MINTS, M., BERTARIO, L., PINEDA, M., NAVARRO, M., MORAK, M., RENKONEN-SINISALO, L., VALENTIN, M. D., FRAYLING, I. M., PLAZZER, J.-P., PYLVANAINEN, K., GENUARDI, M., MECKLIN, J.-P., MOESLEIN, G., SAMPSON, J. R. & CAPELLA, G. 2018. Cancer risk and survival in path_MMR carriers by gene and gender up to 75 years of age: a report from the Prospective Lynch Syndrome Database. *Gut,* 67**,** 1306-1316.
MORI, Y., NAGSE, H., ANDO, H., HORII, A., ICHII, S., NAKATSURU, S., AOKI, T., MIKI, Y., MORI, T. & NAKAMURA, Y. 1992. Somatic mutations of the APC gene in colorectal tumors: mutation cluster region in the APC gene. *Human Molecular Genetics,* 1**,** 229-233.
MORIN, P. J., SPARKS, A. B., KORINEK, V., BARKER, N., CLEVERS, H., VOGELSTEIN, B. & KINZLER, K. W. 1997. Activation of β-Catenin-Tcf Signaling in Colon Cancer by Mutations in β-Catenin or APC. *Science,* 275**,** 1787-1790.
MORIYA, M. & GROLLMAN, A. P. 1993. Mutations in the mutY gene of Escherichia coli enhance the frequency of targeted G:C-->T:a transversions induced by a single 8-oxoguanine residue in single-stranded DNA. *Mol Gen Genet,* 239**,** 72-6.
NAGASE, H., MIYOSHI, Y., HORII, A., AOKI, T., OGAWA, M., UTSUNOMIYA, J., BABA, S., SASAZUKI, T. & NAKAMURA, Y. 1992. Correlation between the Location of Germ-Line Mutations in the *APC* Gene and the Number of Colorectal Polyps in Familial Adenomatous Polyposis Patients. *Cancer Research,* 52**,** 4055-4057.
NAKAGAWA, H., LOCKMAN, J. C., FRANKEL, W. L., HAMPEL, H., STEENBLOCK, K., BURGART, L. J., THIBODEAU, S. N. & DE LA CHAPELLE, A. 2004. Mismatch repair gene PMS2: disease-causing germline mutations are frequent in patients whose tumors stain negative for PMS2 protein, but paralogous genes obscure mutation detection and interpretation. *Cancer Res,* 64**,** 4721-7.
NAKAMURA, Y., LATHROP, M., LEPPERT, M., DOBBS, M., WASMUTH, J., WOLFF, E., CARLSON, M., FUJIMOTO, E., KRAPCHO, K., SEARS, T. & ET AL. 1988. Localization of the genetic defect in familial adenomatous polyposis within a small region of chromosome 5. *Am J Hum Genet,* 43**,** 638-44.
NATIONAL CENTER FOR BIOTECHNOLOGY INFORMATION. *ClinVar; Variation ID 505483* [Online]. Available: https://www.ncbi.nlm.nih.gov/clinvar/variation/505483/ [Accessed 28th August 2019].

NEGRI, G., MILANI, D., COLAPIETRO, P., FORZANO, F., DELLA MONICA, M., RUSCONI, D., CONSONNI, L., CAFFI, L. G., FINELLI, P., SCARANO, G., MAGNANI, C., SELICORNI, A., SPENA, S., LARIZZA, L. & GERVASINI, C. 2015. Clinical and molecular characterization of Rubinstein-Taybi syndrome patients carrying distinct novel mutations of the EP300 gene. *Clinical Genetics,* 87**,** 148-154.

NG, P. C. & HENIKOFF, S. 2003. SIFT: Predicting amino acid changes that affect protein function. *Nucleic Acids Res,* 31**,** 3812-4.

NG, S. B., BIGHAM, A. W., BUCKINGHAM, K. J., HANNIBAL, M. C., MCMILLIN, M. J., GILDERSLEEVE, H. I., BECK, A. E., TABOR, H. K., COOPER, G. M., MEFFORD, H. C., LEE, C., TURNER, E. H., SMITH, J. D., RIEDER, M. J., YOSHIURA, K. I., MATSUMOTO, N., OHTA, T., NIIKAWA, N., NICKERSON, D. A., BAMSHAD, M. J. & SHENDURE, J. 2010a. Exome sequencing identifies MLL2 mutations as a cause of Kabuki syndrome. *Nature Genetics,* 42**,** 790-793.

NG, S. B., BUCKINGHAM, K. J., LEE, C., BIGHAM, A. W., TABOR, H. K., DENT, K. M., HUFF, C. D., SHANNON, P. T., JABS, E. W., NICKERSON, D. A., SHENDURE, J. & BAMSHAD, M. J. 2010b. Exome sequencing identifies the cause of a mendelian disorder. *Nature Genetics,* 42**,** 30-35.

NGEOW, J., YU, W., YEHIA, L., NIAZI, F., CHEN, J., TANG, X., HEALD, B., LEI, J., ROMIGH, T., TUCKER-KELLOGG, L., LIM, K. H., SONG, H. & ENG, C. 2015. Exome Sequencing Reveals Germline SMAD9 Mutation That Reduces Phosphatase and Tensin Homolog Expression and Is Associated With Hamartomatous Polyposis and Gastrointestinal Ganglioneuromas. *Gastroenterology,* 149**,** 886-9 e5.

NICOLAIDES, N. C., PAPADOPOULOS, N., LIU, B., WEIT, Y.-F., CARTER, K. C., RUBEN, S. M., ROSEN, C. A., HASELTINE, W. A., FLEISCHMANN, R. D., FRASER, C. M., ADAMS, M. D., VENTER, J. C., DUNLOP, M. G., HAMILTON, S. R., PETERSEN, G. M., DE LA CHAPELLE, A., VOGELSTEIN, B. & KINZLER, K. W. 1994. Mutations in two PMS homologues in hereditary nonpolyposis colon cancer. *Nature,* 371**,** 75-80.

NIEMINEN, T. T., O'DONOHUE, M. F., WU, Y., LOHI, H., SCHERER, S. W., PATERSON, A. D., ELLONEN, P., ABDEL-RAHMAN, W. M., VALO, S., MECKLIN, J. P., JARVINEN, H. J., GLEIZES, P. E. & PELTOMAKI, P. 2014. Germline mutation of RPS20, encoding a ribosomal protein, causes predisposition to hereditary nonpolyposis colorectal carcinoma without DNA mismatch repair deficiency. *Gastroenterology,* 147**,** 595-598 e5.

OLSEN, J. H., HAHNEMANN, J. M., BORRESEN-DALE, A. L., TRETLI, S., KLEINERMAN, R., SANKILA, R., HAMMARSTROM, L., ROBSAHM, T. E., KAARIAINEN, H., BREGARD, A., BRONDUM-NIELSEN, K., YUEN, J. & TUCKER, M. 2005. Breast and other cancers in 1445 blood relatives of 75 Nordic patients with ataxia telangiectasia. *Br J Cancer,* 93**,** 260-5.

OONO, Y., FU, K., NAKAMURA, H., IRIGUCHI, Y., YAMAMURA, A., TOMINO, Y., ODA, J., MIZUTANI, M., TAKAYANAGI, S., KISHI, D., SHINOHARA, T., YAMADA, K., MATUMOTO, J. & IMAMURA, K. 2009. Progression of a Sessile Serrated Adenoma to an Early Invasive Cancer Within 8 Months. *Digestive Diseases and Sciences,* 54**,** 906-909.

OSBORNE, R. H., HOPPER, J. L., KIRK, J. A., CHENEVIX-TRENCH, G., THORNE, H. J. & SAMBROOK, J. F. 2000. kConFab: a research resource of Australasian breast cancer families. Kathleen Cuningham Foundation Consortium for Research into Familial Breast Cancer. *Med J Aust,* 172**,** 463-4.

PAI, S. I., WU, G. S., ÖZÖREN, N., WU, L., JEN, J., SIDRANSKY, D. & EL-DEIRY, W. S. 1998. Rare Loss-of-Function Mutation of a Death Receptor Gene in Head and Neck Cancer. *Cancer Research,* 58**,** 3513-3518.

PAISÁN-RUíZ, C., JAIN, S., EVANS, E. W., GILKS, W. P., SIMÓN, J., VAN DER BRUG, M., DE MUNAIN, A. L., APARICIO, S., GIL, A. M. N., KHAN, N., JOHNSON, J., MARTINEZ, J. R., NICHOLL, D., CARRERA, I. M., PEÑA, A. S., DE SILVA, R., LEES, A., MARTí-MASSÓ, J. F.,

PÉREZ-TUR, J., WOOD, N. W. & SINGLETON, A. B. 2004. Cloning of the Gene Containing Mutations that Cause PARK8-Linked Parkinson's Disease. *Neuron,* 44**,** 595-600.

PALLES, C., CAZIER, J. B., HOWARTH, K. M., DOMINGO, E., JONES, A. M., BRODERICK, P., KEMP, Z., SPAIN, S. L., GUARINO, E., SALGUERO, I., SHERBORNE, A., CHUBB, D., CARVAJAL-CARMONA, L. G., MA, Y., KAUR, K., DOBBINS, S., BARCLAY, E., GORMAN, M., MARTIN, L., KOVAC, M. B., HUMPHRAY, S., CONSORTIUM, C., CONSORTIUM, W. G. S., LUCASSEN, A., HOLMES, C. C., BENTLEY, D., DONNELLY, P., TAYLOR, J., PETRIDIS, C., ROYLANCE, R., SAWYER, E. J., KERR, D. J., CLARK, S., GRIMES, J., KEARSEY, S. E., THOMAS, H. J., MCVEAN, G., HOULSTON, R. S. & TOMLINSON, I. 2013. Germline mutations affecting the proofreading domains of POLE and POLD1 predispose to colorectal adenomas and carcinomas. *Nat Genet,* 45**,** 136-44.

PALOMBO, F., GALLINARI, P., IACCARINO, I., LETTIERI, T., HUGHES, M., D'ARRIGO, A., TRUONG, O., HSUAN, J. J. & JIRICNY, J. 1995. GTBP, a 160-kilodalton protein essential for mismatch-binding activity in human cells. *Science,* 268**,** 1912-4.

PARK, D. J., LESUEUR, F., NGUYEN-DUMONT, T., PERTESI, M., ODEFREY, F., HAMMET, F., NEUHAUSEN, S. L., JOHN, E. M., ANDRULIS, I. L., TERRY, M. B., DALY, M., BUYS, S., LE CALVEZ-KELM, F., LONIE, A., POPE, B. J., TSIMIKLIS, H., VOEGELE, C., HILBERS, F. M., HOOGERBRUGGE, N., BARROSO, A., OSORIO, A., GILES, G. G., DEVILEE, P., BENITEZ, J., HOPPER, J. L., TAVTIGIAN, S. V., GOLDGAR, D. E. & SOUTHEY, M. C. 2012. Rare Mutations in XRCC2 Increase the Risk of Breast Cancer. *The American Journal of Human Genetics,* 90**,** 734-739.

PAVLOV, Y. I., SHCHERBAKOVA, P. V. & ROGOZIN, I. B. 2006. Roles of DNA Polymerases in Replication, Repair, and Recombination in Eukaryotes. *In:* JEON, K. W. (ed.) *International Review of Cytology.* Academic Press.

PE'ER, I., YELENSKY, R., ALTSHULER, D. & DALY, M. J. 2008. Estimation of the multiple testing burden for genomewide association studies of nearly all common variants. 32**,** 381-385.

PELTOMAKI, P. 2005. Lynch syndrome genes. *Fam Cancer,* 4**,** 227-32.

PELTOMAKI, P., AALTONEN, L., SISTONEN, P., PYLKKANEN, L., MECKLIN, J., JARVINEN, H., GREEN, J., JASS, WEBER, J., LEACH, F. & ET, A. 1993. Genetic mapping of a locus predisposing to human colorectal cancer. *Science,* 260**,** 810-812.

PENNING, T. M., JIN, Y., STECKELBROECK, S., LANIŠNIK RIŽNER, T. & LEWIS, M. 2004. Structure–function of human 3α-hydroxysteroid dehydrogenases: genes and proteins. *Molecular and Cellular Endocrinology,* 215**,** 63-72.

PERRIN-VIDOZ, L., SINILNIKOVA, O. M., STOPPA-LYONNET, D., LENOIR, G. M. & MAZOYER, S. 2002. The nonsense-mediated mRNA decay pathway triggers degradation of most BRCA1 mRNAs bearing premature termination codons. *Human Molecular Genetics,* 11**,** 2805-2814.

PETERS, U., HUTTER, C. M., HSU, L., SCHUMACHER, F. R., CONTI, D. V., CARLSON, C. S., EDLUND, C. K., HAILE, R. W., GALLINGER, S., ZANKE, B. W., LEMIRE, M., RANGREJ, J., VIJAYARAGHAVAN, R., CHAN, A. T., HAZRA, A., HUNTER, D. J., MA, J., FUCHS, C. S., GIOVANNUCCI, E. L., KRAFT, P., LIU, Y., CHEN, L., JIAO, S., MAKAR, K. W., TAVERNA, D., GRUBER, S. B., RENNERT, G., MORENO, V., ULRICH, C. M., WOODS, M. O., GREEN, R. C., PARFREY, P. S., PRENTICE, R. L., KOOPERBERG, C., JACKSON, R. D., LACROIX, A. Z., CAAN, B. J., HAYES, R. B., BERNDT, S. I., CHANOCK, S. J., SCHOEN, R. E., CHANG-CLAUDE, J., HOFFMEISTER, M., BRENNER, H., FRANK, B., BEZIEAU, S., KURY, S., SLATTERY, M. L., HOPPER, J. L., JENKINS, M. A., LE MARCHAND, L., LINDOR, N. M., NEWCOMB, P. A., SEMINARA, D., HUDSON, T. J., DUGGAN, D. J., POTTER, J. D. & CASEY, G. 2012. Meta-analysis of new genome-wide association studies of colorectal cancer risk. *Hum Genet,* 131**,** 217-34.

PETERS, U., JIAO, S., SCHUMACHER, F. R., HUTTER, C. M., ARAGAKI, A. K., BARON, J. A., BERNDT, S. I., BEZIEAU, S., BRENNER, H., BUTTERBACH, K., CAAN, B. J., CAMPBELL, P. T., CARLSON, C. S., CASEY, G., CHAN, A. T., CHANG-CLAUDE, J., CHANOCK, S. J., CHEN, L. S., COETZEE, G. A., COETZEE, S. G., CONTI, D. V., CURTIS, K. R., DUGGAN, D., EDWARDS, T., FUCHS, C. S., GALLINGER, S., GIOVANNUCCI, E. L., GOGARTEN, S. M., GRUBER, S. B., HAILE, R. W., HARRISON, T. A., HAYES, R. B., HENDERSON, B. E., HOFFMEISTER, M., HOPPER, J. L., HUDSON, T. J., HUNTER, D. J., JACKSON, R. D., JEE, S. H., JENKINS, M. A., JIA, W. H., KOLONEL, L. N., KOOPERBERG, C., KURY, S., LACROIX, A. Z., LAURIE, C. C., LAURIE, C. A., LE MARCHAND, L., LEMIRE, M., LEVINE, D., LINDOR, N. M., LIU, Y., MA, J., MAKAR, K. W., MATSUO, K., NEWCOMB, P. A., POTTER, J. D., PRENTICE, R. L., QU, C., ROHAN, T., ROSSE, S. A., SCHOEN, R. E., SEMINARA, D., SHRUBSOLE, M., SHU, X. O., SLATTERY, M. L., TAVERNA, D., THIBODEAU, S. N., ULRICH, C. M., WHITE, E., XIANG, Y., ZANKE, B. W., ZENG, Y. X., ZHANG, B., ZHENG, W. & HSU, L. 2013. Identification of Genetic Susceptibility Loci for Colorectal Tumors in a Genome-Wide Meta-analysis. *Gastroenterology,* 144**,** 799-807 e24.

PETROVSKI, S., WANG, Q., HEINZEN, E. L., ALLEN, A. S. & GOLDSTEIN, D. B. 2013. Genic Intolerance to Functional Variation and the Interpretation of Personal Genomes. *PLOS Genetics,* 9**,** e1003709.

PHELAN, C. M., IQBAL, J., LYNCH, H. T., LUBINSKI, J., GRONWALD, J., MOLLER, P., GHADIRIAN, P., FOULKES, W. D., ARMEL, S., EISEN, A., NEUHAUSEN, S. L., SENTER, L., SINGER, C. F., AINSWORTH, P., KIM-SING, C., TUNG, N., LLACUACHAQUI, M., CHORNOKUR, G., PING, S. & NAROD, S. A. 2014. Incidence of colorectal cancer in BRCA1 and BRCA2 mutation carriers: results from a follow-up study. *Br J Cancer,* 110**,** 530-4.

POPOVA, T., HEBERT, L., JACQUEMIN, V., GAD, S., CAUX-MONCOUTIER, V., DUBOIS-D'ENGHIEN, C., RICHAUDEAU, B., RENAUDIN, X., SELLERS, J., NICOLAS, A., SASTRE-GARAU, X., DESJARDINS, L., GYAPAY, G., RAYNAL, V., SINILNIKOVA, OLGA M., ANDRIEU, N., MANIÉ, E., DE PAUW, A., GESTA, P., BONADONA, V., MAUGARD, CHRISTINE M., PENET, C., AVRIL, M.-F., BARILLOT, E., CABARET, O., DELATTRE, O., RICHARD, S., CARON, O., BENFODDA, M., HU, H.-H., SOUFIR, N., BRESSAC-DE PAILLERETS, B., STOPPA-LYONNET, D. & STERN, M.-H. 2013. Germline BAP1 Mutations Predispose to Renal Cell Carcinomas. *The American Journal of Human Genetics,* 92**,** 974-980.

POWELL , S. M., PETERSEN , G. M., KRUSH , A. J., BOOKER , S., JEN , J., GIARDIELLO , F. M., HAMILTON , S. R., VOGELSTEIN , B. & KINZLER , K. W. 1993. Molecular Diagnosis of Familial Adenomatous Polyposis. *New England Journal of Medicine,* 329**,** 1982-1987.

POWELL, S. M., ZILZ, N., BEAZER-BARCLAY, Y., BRYAN, T. M., HAMILTON, S. R., THIBODEAU, S. N., VOGELSTEIN, B. & KINZLER, K. W. 1992. APC mutations occur early during colorectal tumorigenesis. *Nature,* 359**,** 235-237.

PYATT, R. E., PILARSKI, R. & PRIOR, T. W. 2006. Mutation screening in juvenile polyposis syndrome. *J Mol Diagn,* 8**,** 84-8.

QI, W., TANG, Z. & YU, H. 2006. Phosphorylation- and polo-box-dependent binding of Plk1 to Bub1 is required for the kinetochore localization of Plk1. *Mol Biol Cell,* 17**,** 3705-16.

RACEK, A. R., RABE, K. G., WICK, M. J., PSYCHOGIOS, A. & LINDOR, N. M. 2011. Primary appendiceal mucinous adenocarcinoma in two first-degree relatives: case report and review. *Hered Cancer Clin Pract,* 9**,** 1.

RAMENSKY, V., BORK, P. & SUNYAEV, S. 2002. Human non-synonymous SNPs: server and survey. *Nucleic Acids Res,* 30**,** 3894-900.

RANGANATHAN, P., WEAVER, K. L. & CAPOBIANCO, A. J. 2011. Notch signalling in solid tumours: A little bit of everything but not all the time. *Nature Reviews Cancer,* 11**,** 338-351.

RENKONEN-SINISALO, L., AARNIO, M., MECKLIN, J. P. & JARVINEN, H. J. 2000a. Surveillance improves survival of colorectal cancer in patients with hereditary nonpolyposis colorectal cancer. *Cancer Detection and Prevention,* 24**,** 137-142.

RENKONEN-SINISALO, L., AARNIO, M., MECKLIN, J. P. & JARVINEN, H. J. 2000b. Surveillance improves survival of colorectal cancer in patients with hereditary nonpolyposis colorectal cancer. *Cancer Detect Prev,* 24**,** 137-42.

RIO FRIO, T., LAVOIE, J., HAMEL, N., GEYER, F. C., KUSHNER, Y. B., NOVAK, D. J., WARK, L., CAPELLI, C., REIS-FILHO, J. S., MAI, S., PASTINEN, T., TISCHKOWITZ, M. D., MARCUS, V. A. & FOULKES, W. D. 2010. Homozygous BUB1B mutation and susceptibility to gastrointestinal neoplasia. *N Engl J Med,* 363**,** 2628-37.

RIVERA, B., CASTELLSAGUE, E., BAH, I., VAN KEMPEN, L. C. & FOULKES, W. D. 2015. Biallelic NTHL1 Mutations in a Woman with Multiple Primary Tumors. *N Engl J Med,* 373**,** 1985-6.

ROBERTS, A., NANCARROW, D., CLENDENNING, M., BUCHANAN, D. D., JENKINS, M. A., DUGGAN, D., TAVERNA, D., MCKEONE, D., WALTERS, R., WALSH, M. D., YOUNG, B. W., JASS, J. R., ROSTY, C., GATTAS, M., PELZER, E., HOPPER, J. L., GOLDBLATT, J., GEORGE, J., SUTHERS, G. K., PHILLIPS, K., PARRY, S., WOODALL, S., ARNOLD, J., TUCKER, K., MUIR, A., DRINI, M., MACRAE, F., NEWCOMB, P., POTTER, J. D., PAVLUK, E., LINDBLOM, A. & YOUNG, J. P. 2011. Linkage to chromosome 2q32.2-q33.3 in familial serrated neoplasia (Jass syndrome). *Familial Cancer,* 10**,** 245-254.

ROBERTS, N. J., JIAO, Y., YU, J., KOPELOVICH, L., PETERSEN, G. M., BONDY, M. L., GALLINGER, S., SCHWARTZ, A. G., SYNGAL, S., COTE, M. L., AXILBUND, J., SCHULICK, R., ALI, S. Z., ESHLEMAN, J. R., VELCULESCU, V. E., GOGGINS, M., VOGELSTEIN, B., PAPADOPOULOS, N., HRUBAN, R. H., KINZLER, K. W. & KLEIN, A. P. 2012. ATM mutations in patients with hereditary pancreatic cancer. *Cancer Discovery,* 2**,** 41-46.

ROBERTS, N. J., NORRIS, A. L., PETERSEN, G. M., BONDY, M. L., BRAND, R., GALLINGER, S., KURTZ, R. C., OLSON, S. H., RUSTGI, A. K., SCHWARTZ, A. G., STOFFEL, E., SYNGAL, S., ZOGOPOULOS, G., ALI, S. Z., AXILBUND, J., CHAFFEE, K. G., CHEN, Y. C., COTE, M. L., CHILDS, E. J., DOUVILLE, C., GOES, F. S., HERMAN, J. M., IACOBUZIO-DONAHUE, C., KRAMER, M., MAKOHON-MOORE, A., MCCOMBIE, R. W., MCMAHON, K. W., NIKNAFS, N., PARLA, J., PIROOZNIA, M., POTASH, J. B., RHIM, A. D., SMITH, A. L., WANG, Y., WOLFGANG, C. L., WOOD, L. D., ZANDI, P. P., GOGGINS, M., KARCHIN, R., ESHLEMAN, J. R., PAPADOPOULOS, N., KINZLER, K. W., VOGELSTEIN, B., HRUBAN, R. H. & KLEIN, A. P. 2016. Whole Genome Sequencing Defines the Genetic Heterogeneity of Familial Pancreatic Cancer. *Cancer Discov,* 6**,** 166-75.

ROBINSON, J. T., THORVALDSDÓTTIR, H., WINCKLER, W., GUTTMAN, M., LANDER, E. S., GETZ, G. & MESIROV, J. P. 2011. Integrative genomics viewer. *Nature biotechnology,* 29**,** 24-26.

ROHLIN, A., ZAGORAS, T., NILSSON, S., LUNDSTAM, U., WAHLSTROM, J., HULTEN, L., MARTINSSON, T., KARLSSON, G. B. & NORDLING, M. 2014. A mutation in POLE predisposing to a multi-tumour phenotype. *Int J Oncol,* 45**,** 77-81.

RONNETT, B. M., SHMOOKLER, B. M., DIENER-WEST, M., SUGARBAKER, P. H. & KURMAN, R. J. 1997. Immunohistochemical evidence supporting the appendiceal origin of pseudomyxoma peritonei in women. *Int J Gynecol Pathol,* 16**,** 1-9.

ROZEN, P., SAMUEL, Z., SHOMRAT, R. & LEGUM, C. 1999a. Notable intrafamilial phenotypic variability in a kindred with familial adenomatous polyposis and an APCmutation in exon 9. *Gut,* 45**,** 829-833.

ROZEN, P., SHOMRAT, R., STRUL, H., NAIMAN, T., KARMINSKY, N., LEGUM, C. & ORR-URTREGER, A. 1999b. Prevalence of the I1307K APC gene variant in Israeli Jews of differing ethnic origin and risk for colorectal cancer. *Gastroenterology,* 116**,** 54-57.

RYLAND, G. L., HUNTER, S. M., DOYLE, M. A., CARAMIA, F., LI, J., ROWLEY, S. M., CHRISTIE, M., ALLAN, P. E., STEPHENS, A. N., BOWTELL, D. D. L., CAMPBELL, I. G. & GORRINGE, K. L. 2015. Mutational landscape of mucinous ovarian carcinoma and its neoplastic precursors. *Genome Medicine,* 7**,** 87.

SAARINEN, S., AAVIKKO, M., AITTOMÄKI, K., LAUNONEN, V., LEHTONEN, R., FRANSSILA, K., LEHTONEN, H. J., KAASINEN, E., BRODERICK, P., TARKKANEN, J., BAIN, B. J., BAUDUER, F., ÜNAL, A., SWERDLOW, A. J., COOKE, R., MÄKINEN, M. J., HOULSTON, R., VAHTERISTO, P. & AALTONEN, L. A. 2011. Exome sequencing reveals germline NPAT mutation as a candidate risk factor for Hodgkin lymphoma. *Blood,* 118**,** 493-498.

SAMOCHA, K. E., ROBINSON, E. B., SANDERS, S. J., STEVENS, C., SABO, A., MCGRATH, L. M., KOSMICKI, J. A., REHNSTROM, K., MALLICK, S., KIRBY, A., WALL, D. P., MACARTHUR, D. G., GABRIEL, S. B., DEPRISTO, M., PURCELL, S. M., PALOTIE, A., BOERWINKLE, E., BUXBAUM, J. D., COOK, E. H., JR., GIBBS, R. A., SCHELLENBERG, G. D., SUTCLIFFE, J. S., DEVLIN, B., ROEDER, K., NEALE, B. M. & DALY, M. J. 2014. A framework for the interpretation of de novo mutation in human disease. *Nat Genet,* 46**,** 944-50.

SAYED, M. G., AHMED, A. F., RINGOLD, J. R., ANDERSON, M. E., BAIR, J. L., MITROS, F. A., LYNCH, H. T., TINLEY, S. T., PETERSEN, G. M., GIARDIELLO, F. M., VOGELSTEIN, B. & HOWE, J. R. 2002. GermlineSMAD4 orBMPRIA mutations and phenotype of juvenile polyposis. *Annals of Surgical Oncology,* 9**,** 901-906.

SCHULTHEIS, A. M., NGUYEN, G. P., ORTMANN, M., KRUIS, W., #XFC, TTNER, R., SCHILDHAUS, H.-U. & MARKIEFKA, B. 2014. Squamous Cell Carcinoma of the Pancreas in a Patient with Germline BRCA2 Mutation-Response to Neoadjuvant Radiochemotherapy. *Case Reports in Oncological Medicine,* 2014**,** 5.

SCHUMACHER, F. R., SCHMIT, S. L., JIAO, S., EDLUND, C. K., WANG, H., ZHANG, B., HSU, L., HUANG, S. C., FISCHER, C. P., HARJU, J. F., IDOS, G. E., LEJBKOWICZ, F., MANION, F. J., MCDONNELL, K., MCNEIL, C. E., MELAS, M., RENNERT, H. S., SHI, W., THOMAS, D. C., VAN DEN BERG, D. J., HUTTER, C. M., ARAGAKI, A. K., BUTTERBACH, K., CAAN, B. J., CARLSON, C. S., CHANOCK, S. J., CURTIS, K. R., FUCHS, C. S., GALA, M., GIOVANNUCCI, E. L., GOGARTEN, S. M., HAYES, R. B., HENDERSON, B., HUNTER, D. J., JACKSON, R. D., KOLONEL, L. N., KOOPERBERG, C., KURY, S., LACROIX, A., LAURIE, C. C., LAURIE, C. A., LEMIRE, M., LEVINE, D., MA, J., MAKAR, K. W., QU, C., TAVERNA, D., ULRICH, C. M., WU, K., KONO, S., WEST, D. W., BERNDT, S. I., BEZIEAU, S., BRENNER, H., CAMPBELL, P. T., CHAN, A. T., CHANG-CLAUDE, J., COETZEE, G. A., CONTI, D. V., DUGGAN, D., FIGUEIREDO, J. C., FORTINI, B. K., GALLINGER, S. J., GAUDERMAN, W. J., GILES, G., GREEN, R., HAILE, R., HARRISON, T. A., HOFFMEISTER, M., HOPPER, J. L., HUDSON, T. J., JACOBS, E., IWASAKI, M., JEE, S. H., JENKINS, M., JIA, W. H., JOSHI, A., LI, L., LINDOR, N. M., MATSUO, K., MORENO, V., MUKHERJEE, B., NEWCOMB, P. A., POTTER, J. D., RASKIN, L., RENNERT, G., ROSSE, S., SEVERI, G., SCHOEN, R. E., SEMINARA, D., SHU, X. O., SLATTERY, M. L., TSUGANE, S., WHITE, E., XIANG, Y. B., ZANKE, B. W., ZHENG, W., LE MARCHAND, L., CASEY, G., GRUBER, S. B., et al. 2015. Genome-wide association study of colorectal cancer identifies six new susceptibility loci. *Nat Commun,* 6**,** 7138.

SCOTT, S. P., BENDIX, R., CHEN, P., CLARK, R., DÖRK, T. & LAVIN, M. F. 2002. Missense mutations but not allelic variants alter the function of ATM by dominant interference in patients with breast cancer. *Proceedings of the National Academy of Sciences,* 99**,** 925-930.

SEGDITSAS, S. & TOMLINSON, I. 2006. Colorectal cancer and genetic alterations in the Wnt pathway. *Oncogene,* 25**,** 7531-7537.

SEGUI, N., MINA, L. B., LAZARO, C., SANZ-PAMPLONA, R., PONS, T., NAVARRO, M., BELLIDO, F., LOPEZ-DORIGA, A., VALDES-MAS, R., PINEDA, M., GUINO, E., VIDAL, A., SOTO, J. L., CALDES, T., DURAN, M., URIOSTE, M., RUEDA, D., BRUNET, J., BALBIN, M., BLAY, P., IGLESIAS, S., GARRE, P., LASTRA, E., SANCHEZ-HERAS, A. B., VALENCIA, A., MORENO, V.,

PUJANA, M. A., VILLANUEVA, A., BLANCO, I., CAPELLA, G., SURRALLES, J., PUENTE, X. S. & VALLE, L. 2015. Germline Mutations in FAN1 Cause Hereditary Colorectal Cancer by Impairing DNA Repair. *Gastroenterology,* 149**,** 563-6.

SHAIB, W. L., GOODMAN, M., CHEN, Z., KIM, S., BRUTCHER, E., BEKAII-SAAB, T. & EL-RAYES, B. F. 2017. Incidence and Survival of Appendiceal Mucinous Neoplasms: A SEER Analysis. *American Journal of Clinical Oncology,* 40**,** 569-573.

SHEN, T., LEE, A., SHEN, C. & LIN, C. J. 2015. The long tail and rare disease research: the impact of next-generation sequencing for rare Mendelian disorders. *Genetics Research,* 97.

SHERIDAN, J. P., MARSTERS, S. A., PITTI, R. M. & ASHKENAZI, A. 1997. Control of TRAIL-induced apoptosis by a family of signaling and decoy receptors. *Science,* 277**,** 818+.

SHERRY, S. T., WARD, M. H., KHOLODOV, M., BAKER, J., PHAN, L., SMIGIELSKI, E. M. & SIROTKIN, K. 2001. dbSNP: the NCBI database of genetic variation. *Nucleic Acids Research,* 29**,** 308-311.

SHIH, I. M., YAN, H., SPEYRER, D., SHMOOKLER, B. M., SUGARBAKER, P. H. & RONNETT, B. M. 2001. Molecular genetic analysis of appendiceal mucinous adenomas in identical twins, including one with pseudomyxoma peritonei. *Am J Surg Pathol,* 25**,** 1095-9.

SHTOYERMAN-CHEN, R., FRIEDMAN, E., FIGER, A., CARMEL, M., PATAEL, Y., RATH, P., FIDDER, H. H., BAR-MEIR, S. & THEODOR, L. 2001. The I1307K APC polymorphism: prevalence in non-Ashkenazi Jews and evidence for a founder effect. *Genet Test,* 5**,** 141-6.

SIEBER, O. M., LIPTON, L., CRABTREE, M., HEINIMANN, K., FIDALGO, P., PHILLIPS, R. K. S., BISGAARD, M. L., ORNTOFT, T. F., AALTONEN, L. A., HODGSON, S. V., THOMAS, H. J. W. & TOMLINSON, I. P. M. 2003. Multiple colorectal adenomas, classic adenomatous polyposis, and germ-line mutations in MYH. *New England Journal of Medicine,* 348**,** 791-799.

SIERRA POTCHANANT, E. A., CERABONA, D., SATER, Z. A., HE, Y., SUN, Z., GEHLHAUSEN, J. & NALEPA, G. 2017a. INPP5E Preserves Genomic Stability through Regulation of Mitosis. *Molecular and Cellular Biology,* 37.

SIERRA POTCHANANT, E. A., CERABONA, D., SATER, Z. A., HE, Y., SUN, Z., GEHLHAUSEN, J. & NALEPA, G. 2017b. INPP5E Preserves Genomic Stability through Regulation of Mitosis. *Mol Cell Biol,* 37.

SJURSEN, W., MCPHILLIPS, M., SCOTT, R. J. & TALSETH-PALMER, B. A. 2016. Lynch syndrome mutation spectrum in New South Wales, Australia, including 55 novel mutations. *Molecular genetics & genomic medicine,* 4**,** 223-231.

SLOANE, M. A., HESSON, L. B., NUNEZ, A. C., THOMPSON, B. A. & WARD, R. L. 2014. Nucleosome positioning is unaltered at MLH1 splice site mutations in cells derived from Lynch syndrome patients. *Clinical epigenetics,* 6**,** 32-32.

SMITH, C. G., NAVEN, M., HARRIS, R., COLLEY, J., WEST, H., LI, N., LIU, Y., ADAMS, R., MAUGHAN, T. S., NICHOLS, L., KAPLAN, R., WAGNER, M. J., MCLEOD, H. L. & CHEADLE, J. P. 2013. Exome Resequencing Identifies Potential Tumor-Suppressor Genes that Predispose to Colorectal Cancer. *Hum Mutat,* 34**,** 1026-34.

SNOVER, D., AHNEN, D., BURT, R. & ODZE, R. 2010. Serrated polyps of the colon and rectum and serrated ("hyperplastic") polyposis. *In:* BOSMAN S.T., C. F., HRUBAN R.H. , THEISE N.D. (ed.) *WHO Classification of tumours of the digestive system.* Berlin: Springer-Verlag.

SOPIK, V., PHELAN, C., CYBULSKI, C. & NAROD, S. A. 2015. BRCA1 and BRCA2 mutations and the risk for colorectal cancer. *Clin Genet,* 87**,** 411-8.

STADLER, Z. K., BATTAGLIN, F., MIDDHA, S., HECHTMAN, J. F., TRAN, C., CERCEK, A., YAEGER, R., SEGAL, N. H., VARGHESE, A. M., REIDY-LAGUNES, D. L., KEMENY, N. E., SALO-MULLEN, E. E., ASHRAF, A., WEISER, M. R., GARCIA-AGUILAR, J., ROBSON, M. E., OFFIT, K., ARCILA, M. E., BERGER, M. F., SHIA, J., SOLIT, D. B. & SALTZ, L. B. 2016. Reliable

Detection of Mismatch Repair Deficiency in Colorectal Cancers Using Mutational Load in Next-Generation Sequencing Panels. *J Clin Oncol,* 34**,** 2141-7.

STRAND, M., PROLLA, T. A., LISKAY, R. M. & PETES, T. D. 1993. Destabilization of tracts of simple repetitive DNA in yeast by mutations affecting DNA mismatch repair. *Nature,* 365**,** 274-6.

SUDAKIN, V., CHAN, G. K. T. & YEN, T. J. 2001. Checkpoint inhibition of the APC/C in HeLa cells is mediated by a complex of BUBR1, BUB3, CDC20, and MAD2. *The Journal of Cell Biology,* 154**,** 925-936.

TAGGART, M. W., GALBINCEA, J., MANSFIELD, P. F., FOURNIER, K. F., ROYAL, R. E., OVERMAN, M. J., RASHID, A. & ABRAHAM, S. C. 2013. High-level microsatellite instability in appendiceal carcinomas. *American Journal of Surgical Pathology,* 37**,** 1192-1200.

TANAKA, M., KUSUMI, T., SASAKI, Y., YAMAGATA, K., ICHINOHE, H., NISHIDA, J. & KUDO, H. 2001. Colonic intra-epithelial carcinoma occurring in a hyperplastic polyp via a serrated adenoma. *Pathology International,* 51**,** 215-220.

TANSKANEN, T., GYLFE, A. E., KATAINEN, R., TAIPALE, M., RENKONEN-SINISALO, L., JARVINEN, H., MECKLIN, J. P., BOHM, J., KILPIVAARA, O., PITKANEN, E., PALIN, K., VAHTERISTO, P., TUUPANEN, S. & AALTONEN, L. A. 2015. Systematic search for rare variants in Finnish early-onset colorectal cancer patients. *Cancer Genet,* 208**,** 35-40.

TAUPIN, D., LAM, W., RANGIAH, D., MCCALLUM, L., WHITTLE, B., ZHANG, Y., ANDREWS, D., FIELD, M., GOODNOW, C. C. & COOK, M. C. 2015. A deleterious RNF43 germline mutation in a severely affected serrated polyposis kindred. *Human Genome Variation,* 2**,** 15013.

TENESA, A. & DUNLOP, M. G. 2009. New insights into the aetiology of colorectal cancer from genome-wide association studies. *Nat Rev Genet,* 10**,** 353-358.

TENESA, A., FARRINGTON, S. M., PRENDERGAST, J. G., PORTEOUS, M. E., WALKER, M., HAQ, N., BARNETSON, R. A., THEODORATOU, E., CETNARSKYJ, R., CARTWRIGHT, N., SEMPLE, C., CLARK, A. J., REID, F. J., SMITH, L. A., KAVOUSSANAKIS, K., KOESSLER, T., PHAROAH, P. D., BUCH, S., SCHAFMAYER, C., TEPEL, J., SCHREIBER, S., VOLZKE, H., SCHMIDT, C. O., HAMPE, J., CHANG-CLAUDE, J., HOFFMEISTER, M., BRENNER, H., WILKENING, S., CANZIAN, F., CAPELLA, G., MORENO, V., DEARY, I. J., STARR, J. M., TOMLINSON, I. P., KEMP, Z., HOWARTH, K., CARVAJAL-CARMONA, L., WEBB, E., BRODERICK, P., VIJAYAKRISHNAN, J., HOULSTON, R. S., RENNERT, G., BALLINGER, D., ROZEK, L., GRUBER, S. B., MATSUDA, K., KIDOKORO, T., NAKAMURA, Y., ZANKE, B. W., GREENWOOD, C. M., RANGREJ, J., KUSTRA, R., MONTPETIT, A., HUDSON, T. J., GALLINGER, S., CAMPBELL, H. & DUNLOP, M. G. 2008. Genome-wide association scan identifies a colorectal cancer susceptibility locus on 11q23 and replicates risk loci at 8q24 and 18q21. *Nat Genet,* 40**,** 631-7.

TERDIMAN, J. P. 2009. MYH-Associated Disease: Attenuated Adenomatous Polyposis of the Colon Is Only Part of the Story. *Gastroenterology,* 137**,** 1883-1886.

THE CANCER GENOME ATLAS, N. 2012. Comprehensive Molecular Characterization of Human Colon and Rectal Cancer. *Nature,* 487**,** 330-337.

THE ENCODE PROJECT CONSORTIUM, DUNHAM, I., KUNDAJE, A., ALDRED, S. F., COLLINS, P. J., DAVIS, C. A., DOYLE, F., EPSTEIN, C. B., FRIETZE, S., HARROW, J., KAUL, R., KHATUN, J., LAJOIE, B. R., LANDT, S. G., LEE, B.-K., PAULI, F., ROSENBLOOM, K. R., SABO, P., SAFI, A., SANYAL, A., SHORESH, N., SIMON, J. M., SONG, L., TRINKLEIN, N. D., ALTSHULER, R. C., BIRNEY, E., BROWN, J. B., CHENG, C., DJEBALI, S., DONG, X., DUNHAM, I., ERNST, J., FUREY, T. S., GERSTEIN, M., GIARDINE, B., GREVEN, M., HARDISON, R. C., HARRIS, R. S., HERRERO, J., HOFFMAN, M. M., IYER, S., KELLIS, M., KHATUN, J., KHERADPOUR, P., KUNDAJE, A., LASSMANN, T., LI, Q., LIN, X., MARINOV, G. K., MERKEL, A., MORTAZAVI, A., PARKER, S. C. J., REDDY, T. E., ROZOWSKY, J., SCHLESINGER, F., THURMAN, R. E., WANG, J., WARD, L. D., WHITFIELD, T. W., WILDER, S. P., WU, W., XI, H. S., YIP, K. Y.,

ZHUANG, J., BERNSTEIN, B. E., BIRNEY, E., DUNHAM, I., GREEN, E. D., GUNTER, C., SNYDER, M., PAZIN, M. J., LOWDON, R. F., DILLON, L. A. L., ADAMS, L. B., KELLY, C. J., ZHANG, J., WEXLER, J. R., GREEN, E. D., GOOD, P. J., FEINGOLD, E. A., BERNSTEIN, B. E., BIRNEY, E., CRAWFORD, G. E., DEKKER, J., ELNITSKI, L., FARNHAM, P. J., GERSTEIN, M., GIDDINGS, M. C., GINGERAS, T. R., GREEN, E. D., GUIGÓ, R., HARDISON, R. C., HUBBARD, T. J., KELLIS, M., KENT, W. J., LIEB, J. D., MARGULIES, E. H., MYERS, R. M., SNYDER, M., STAMATOYANNOPOULOS, J. A., et al. 2012. An integrated encyclopedia of DNA elements in the human genome. *Nature,* 489**,** 57.

THE GENE ONTOLOGY CONSORTIUM 2015. Gene Ontology Consortium: going forward. *Nucleic Acids Research,* 43**,** D1049-D1056.

THE UNIPROT CONSORTIUM 2018. UniProt: a worldwide hub of protein knowledge. *Nucleic Acids Research,* 47**,** D506-D515.

THOMAS, D., SCOT, A. D., BARBEY, R., PADULA, M. & BOITEUX, S. 1997. Inactivation of OGG1 increases the incidence of G . C-->T . A transversions in Saccharomyces cerevisiae: evidence for endogenous oxidative damage to DNA in eukaryotic cells. *Mol Gen Genet,* 254**,** 171-8.

THOMPSON, D., DUEDAL, S., KIRNER, J., MCGUFFOG, L., LAST, J., REIMAN, A., BYRD, P., TAYLOR, M. & EASTON, D. F. 2005. Cancer risks and mortality in heterozygous ATM mutation carriers. *J Natl Cancer Inst,* 97**,** 813-22.

THOMPSON, E. R., DOYLE, M. A., RYLAND, G. L., ROWLEY, S. M., CHOONG, D. Y. H., TOTHILL, R. W., THORNE, H., BARNES, D. R., LI, J., ELLUL, J., PHILIP, G. K., ANTILL, Y. C., JAMES, P. A., TRAINER, A. H., MITCHELL, G. & CAMPBELL, I. G. 2012. Exome Sequencing Identifies Rare Deleterious Mutations in DNA Repair Genes FANCC and BLM as Potential Breast Cancer Susceptibility Alleles. *PLoS Genetics,* 8.

TOMLINSON, I., WEBB, E., CARVAJAL-CARMONA, L., BRODERICK, P., KEMP, Z., SPAIN, S., PENEGAR, S., CHANDLER, I., GORMAN, M., WOOD, W., BARCLAY, E., LUBBE, S., MARTIN, L., SELLICK, G., JAEGER, E., HUBNER, R., WILD, R., ROWAN, A., FIELDING, S., HOWARTH, K., SILVER, A., ATKIN, W., MUIR, K., LOGAN, R., KERR, D., JOHNSTONE, E., SIEBER, O., GRAY, R., THOMAS, H., PETO, J., CAZIER, J. B. & HOULSTON, R. 2007. A genome-wide association scan of tag SNPs identifies a susceptibility variant for colorectal cancer at 8q24.21. *Nat Genet,* 39**,** 984-8.

TOMLINSON, I. P., CARVAJAL-CARMONA, L. G., DOBBINS, S. E., TENESA, A., JONES, A. M., HOWARTH, K., PALLES, C., BRODERICK, P., JAEGER, E. E., FARRINGTON, S., LEWIS, A., PRENDERGAST, J. G., PITTMAN, A. M., THEODORATOU, E., OLVER, B., WALKER, M., PENEGAR, S., BARCLAY, E., WHIFFIN, N., MARTIN, L., BALLEREAU, S., LLOYD, A., GORMAN, M., LUBBE, S., HOWIE, B., MARCHINI, J., RUIZ-PONTE, C., FERNANDEZ-ROZADILLA, C., CASTELLS, A., CARRACEDO, A., CASTELLVI-BEL, S., DUGGAN, D., CONTI, D., CAZIER, J. B., CAMPBELL, H., SIEBER, O., LIPTON, L., GIBBS, P., MARTIN, N. G., MONTGOMERY, G. W., YOUNG, J., BAIRD, P. N., GALLINGER, S., NEWCOMB, P., HOPPER, J., JENKINS, M. A., AALTONEN, L. A., KERR, D. J., CHEADLE, J., PHAROAH, P., CASEY, G., HOULSTON, R. S. & DUNLOP, M. G. 2011. Multiple common susceptibility variants near BMP pathway loci GREM1, BMP4, and BMP2 explain part of the missing heritability of colorectal cancer. *PLoS Genet,* 7**,** e1002105.

TOMLINSON, I. P., WEBB, E., CARVAJAL-CARMONA, L., BRODERICK, P., HOWARTH, K., PITTMAN, A. M., SPAIN, S., LUBBE, S., WALTHER, A., SULLIVAN, K., JAEGER, E., FIELDING, S., ROWAN, A., VIJAYAKRISHNAN, J., DOMINGO, E., CHANDLER, I., KEMP, Z., QURESHI, M., FARRINGTON, S. M., TENESA, A., PRENDERGAST, J. G., BARNETSON, R. A., PENEGAR, S., BARCLAY, E., WOOD, W., MARTIN, L., GORMAN, M., THOMAS, H., PETO, J., BISHOP, D. T., GRAY, R., MAHER, E. R., LUCASSEN, A., KERR, D., EVANS, D. G., SCHAFMAYER, C., BUCH, S., VOLZKE, H., HAMPE, J., SCHREIBER, S., JOHN, U., KOESSLER, T., PHAROAH, P., VAN WEZEL, T., MORREAU, H., WIJNEN, J. T., HOPPER, J.

L., SOUTHEY, M. C., GILES, G. G., SEVERI, G., CASTELLVI-BEL, S., RUIZ-PONTE, C., CARRACEDO, A., CASTELLS, A., FORSTI, A., HEMMINKI, K., VODICKA, P., NACCARATI, A., LIPTON, L., HO, J. W., CHENG, K. K., SHAM, P. C., LUK, J., AGUNDEZ, J. A., LADERO, J. M., DE LA HOYA, M., CALDES, T., NIITTYMAKI, I., TUUPANEN, S., KARHU, A., AALTONEN, L., CAZIER, J. B., CAMPBELL, H., DUNLOP, M. G. & HOULSTON, R. S. 2008. A genome-wide association study identifies colorectal cancer susceptibility loci on chromosomes 10p14 and 8q23.3. *Nat Genet,* 40**,** 623-30.

TOYOTA, M., AHUJA, N., OHE-TOYOTA, M., HERMAN, J. G., BAYLIN, S. B. & ISSA, J. P. 1999. CpG island methylator phenotype in colorectal cancer. *Proc Natl Acad Sci U S A,* 96**,** 8681-6.

TRAINER, A. H. 2009. Extra-colonic manifestations of familial adenomatous polyposis coli. *Adv Exp Med Biol,* 656**,** 119-27.

VALLE, L., HERNÁNDEZ-ILLÁN, E., BELLIDO, F., AIZA, G., CASTILLEJO, A., CASTILLEJO, M. I., NAVARRO, M., SEGUÍ, N., VARGAS, G., GUARINOS, C., JUAREZ, M., SANJUÁN, X., IGLESIAS, S., ALENDA, C., EGOAVIL, C., SEGURA, A., JUAN, M. J., RODRIGUEZ-SOLER, M., BRUNET, J., GONZÁLEZ, S., JOVER, R., LÁZARO, C., CAPELLÁ, G., PINEDA, M., SOTO, J. L. & BLANCO, I. 2014. New insights into POLE and POLD1 germline mutations in familial colorectal cancer and polyposis. *Human Molecular Genetics,* 23**,** 3506-3512.

VAN DER KLIFT, H. M., MENSENKAMP, A. R., DROST, M., BIK, E. C., VOS, Y. J., GILLE, H. J. J. P., REDEKER, B. E. J. W., TIERSMA, Y., ZONNEVELD, J. B. M., GARCÍA, E. G., LETTEBOER, T. G. W., OLDERODE-BERENDS, M. J. W., VAN HEST, L. P., VAN OS, T. A., VERHOEF, S., WAGNER, A., VAN ASPEREN, C. J., TEN BROEKE, S. W., HES, F. J., DE WIND, N., NIELSEN, M., DEVILEE, P., LIGTENBERG, M. J. L., WIJNEN, J. T. & TOPS, C. M. J. 2016. Comprehensive Mutation Analysis of PMS2 in a Large Cohort of Probands Suspected of Lynch Syndrome or Constitutional Mismatch Repair Deficiency Syndrome. *Human Mutation,* 37**,** 1162-1179.

VAN DER VELDE, K. J., KUIPER, J., THOMPSON, B. A., PLAZZER, J.-P., VAN VALKENHOEF, G., DE HAAN, M., JONGBLOED, J. D. H., WIJMENGA, C., DE KONING, T. J., ABBOTT, K. M., SINKE, R., SPURDLE, A. B., MACRAE, F., GENUARDI, M., SIJMONS, R. H., SWERTZ, M. A. & INSI, G. H. T. G. 2015. Evaluation of CADD Scores in Curated Mismatch Repair Gene Variants Yields a Model for Clinical Validation and Prioritization. *Human Mutation,* 36**,** 712-719.

VAN OS, N. J., ROELEVELD, N., WEEMAES, C. M., JONGMANS, M. C., JANSSENS, G. O., TAYLOR, A. M., HOOGERBRUGGE, N. & WILLEMSEN, M. A. 2016. Health risks for ataxia-telangiectasia mutated heterozygotes: a systematic review, meta-analysis and evidence-based guideline. *Clin Genet,* 90**,** 105-17.

VASEN, H. F., MECKLIN, J. P., KHAN, P. M. & LYNCH, H. T. 1991. The International Collaborative Group on Hereditary Non-Polyposis Colorectal Cancer (ICG-HNPCC). *Dis Colon Rectum,* 34**,** 424-5.

VINSON, K. E., GEORGE, D. C., FENDER, A. W., BERTRAND, F. E. & SIGOUNAS, G. 2016. The Notch pathway in colorectal cancer. *Int J Cancer,* 138**,** 1835-42.

VOGELSTEIN , B., FEARON , E. R., HAMILTON , S. R., KERN , S. E., PREISINGER , A. C., LEPPERT , M., SMITS , A. M. M. & BOS , J. L. 1988. Genetic Alterations during Colorectal-Tumor Development. *New England Journal of Medicine,* 319**,** 525-532.

VOGT, S., JONES, N., CHRISTIAN, D., ENGEL, C., NIELSEN, M., KAUFMANN, A., STEINKE, V., VASEN, H. F., PROPPING, P., SAMPSON, J. R., HES, F. J. & ARETZ, S. 2009. Expanded Extracolonic Tumor Spectrum in MUTYH-Associated Polyposis. *Gastroenterology,* 137**,** 1976-1985.e10.

WALLBERG, A. E., PEDERSEN, K., LENDAHL, U. & ROEDER, R. G. 2002. p300 and PCAF Act Cooperatively To Mediate Transcriptional Activation from Chromatin Templates by Notch Intracellular Domains In Vitro. *Molecular and Cellular Biology,* 22**,** 7812-7819.

WANG, H., BURNETT, T., KONO, S., HAIMAN, C. A., IWASAKI, M., WILKENS, L. R., LOO, L. W., VAN DEN BERG, D., KOLONEL, L. N., HENDERSON, B. E., KEKU, T. O., SANDLER, R. S., SIGNORELLO, L. B., BLOT, W. J., NEWCOMB, P. A., PANDE, M., AMOS, C. I., WEST, D. W., BEZIEAU, S., BERNDT, S. I., ZANKE, B. W., HSU, L., LINDOR, N. M., HAILE, R. W., HOPPER, J. L., JENKINS, M. A., GALLINGER, S., CASEY, G., STENZEL, S. L., SCHUMACHER, F. R., PETERS, U., GRUBER, S. B., TSUGANE, S., STRAM, D. O. & LE MARCHAND, L. 2014. Trans-ethnic genome-wide association study of colorectal cancer identifies a new susceptibility locus in VTI1A. *Nat Commun,* 5**,** 4613.

WANG, Y., COTTMAN, M. & SCHIFFMAN, J. D. 2012. Molecular inversion probes: a novel microarray technology and its application in cancer research. *Cancer Genet,* 205**,** 341-55.

WARD, R. & HAWKINS, N. 2013. *Molecular and Cellular Oncology (MCO) study tumour collection* [Online]. University of New South Wales. Available: http://dx.doi.org/10.4225/53/5559205bea135, https://researchdata.ands.org.au/mco-study-tumour-collection [Accessed].

WARTHIN, A. 1913. Classics in oncology. Heredity with reference to carcinoma as shown by the study of the cases examined in the pathological laboratory of the University of Michigan, 1895-1913. By Aldred Scott Warthin. 1913. *CA Cancer J Clin,* 35**,** 348-59.

WATSON, P., VASEN, H. F. A., MECKLIN, J.-P., BERNSTEIN, I., AARNIO, M., JÄRVINEN, H. J., MYRHØJ, T., SUNDE, L., WIJNEN, J. T. & LYNCH, H. T. 2008a. The risk of extra-colonic, extra-endometrial cancer in the Lynch syndrome. *International Journal of Cancer,* 123**,** 444-449.

WATSON, P., VASEN, H. F. A., MECKLIN, J. P., BERNSTEIN, I., AARNIO, M., JARVINEN, H. J., MYRHOJ, T., SUNDE, L., WIJNEN, J. T. & LYNCH, H. T. 2008b. The risk of extra-colonic, extra-endometrial cancer in the Lynch syndrome. *Int J Cancer,* 123**,** 444-449.

WEREN, R. D., LIGTENBERG, M. J., KETS, C. M., DE VOER, R. M., VERWIEL, E. T., SPRUIJT, L., VAN ZELST-STAMS, W. A., JONGMANS, M. C., GILISSEN, C., HEHIR-KWA, J. Y., HOISCHEN, A., SHENDURE, J., BOYLE, E. A., KAMPING, E. J., NAGTEGAAL, I. D., TOPS, B. B., NAGENGAST, F. M., GEURTS VAN KESSEL, A., VAN KRIEKEN, J. H., KUIPER, R. P. & HOOGERBRUGGE, N. 2015. A germline homozygous mutation in the base-excision repair gene NTHL1 causes adenomatous polyposis and colorectal cancer. *Nat Genet,* 47**,** 668-71.

WHIFFIN, N., HOSKING, F. J., FARRINGTON, S. M., PALLES, C., DOBBINS, S. E., ZGAGA, L., LLOYD, A., KINNERSLEY, B., GORMAN, M., TENESA, A., BRODERICK, P., WANG, Y., BARCLAY, E., HAYWARD, C., MARTIN, L., BUCHANAN, D. D., WIN, A. K., HOPPER, J., JENKINS, M., LINDOR, N. M., NEWCOMB, P. A., GALLINGER, S., CONTI, D., SCHUMACHER, F., CASEY, G., LIU, T., CAMPBELL, H., LINDBLOM, A., HOULSTON, R. S., TOMLINSON, I. P. & DUNLOP, M. G. 2014. Identification of susceptibility loci for colorectal cancer in a genome-wide meta-analysis. *Hum Mol Genet,* 23**,** 4729-37.

WHITE, B. D., CHIEN, A. J. & DAWSON, D. W. 2012. Dysregulation of Wnt/β-catenin Signaling in Gastrointestinal Cancers. *Gastroenterology,* 142**,** 219-232.

WHITELAW, S. C., MURDAY, V. A., TOMLINSON, I. P., THOMAS, H. J., COTTRELL, S., GINSBERG, A., BUKOFZER, S., HODGSON, S. V., SKUDOWITZ, R. B., JASS, J. R., TALBOT, I. C., NORTHOVER, J. M., BODMER, W. F. & SOLOMON, E. 1997. Clinical and molecular features of the hereditary mixed polyposis syndrome. *Gastroenterology,* 112**,** 327-334.

WIJNEN , J. T., VASEN , H. F. A., KHAN , P. M., ZWINDERMAN , A. H., VAN DER KLIFT , H., MULDER , A., TOPS , C., MØLLER , P., FODDE , R., MENKO , F., TAAL , B., NAGENGAST , F., BRUNNER , H., KLEIBEUKER , J., SIJMONS , R., GRIFFIOEN , G., BRÖCKER-VRIENDS , A., BAKKER , E., VAN LEEUWEN-CORNELISSE , I., MEIJERS-HEIJBOER , A., LINDHOUT , D., BREUNING , M., POST , J., SCHAAP , C., APOLD , J., HEIMDAL , K., BERTARIO , L., BISGAARD , M. L. & GOETZ , P. 1998. Clinical Findings with Implications for Genetic

Testing in Families with Clustering of Colorectal Cancer. *New England Journal of Medicine,* 339**,** 511-518.

WIN, A. K., DOWTY, J. G., CLEARY, S. P., KIM, H., BUCHANAN, D. D., YOUNG, J. P., CLENDENNING, M., ROSTY, C., MACINNIS, R. J., GILES, G. G., BOUSSIOUTAS, A., MACRAE, F. A., PARRY, S., GOLDBLATT, J., BARON, J. A., BURNETT, T., LE MARCHAND, L., NEWCOMB, P. A., HAILE, R. W., HOPPER, J. L., COTTERCHIO, M., GALLINGER, S., LINDOR, N. M., TUCKER, K. M., WINSHIP, I. M. & JENKINS, M. A. 2014. Risk of Colorectal Cancer for Carriers of Mutations in MUTYH, With and Without a Family History of Cancer. *Gastroenterology,* 146**,** 1208-1211.e5.

WIN, A. K., WALTERS, R. J., BUCHANAN, D. D., JENKINS, M. A., SWEET, K., FRANKEL, W. L., DE LA CHAPELLE, A., MCKEONE, D. M., WALSH, M. D., CLENDENNING, M., PEARSON, S.-A., PAVLUK, E., NAGLER, B., HOPPER, J. L., GATTAS, M. R., GOLDBLATT, J., GEORGE, J., SUTHERS, G. K., PHILLIPS, K. D., WOODALL, S., ARNOLD, J., TUCKER, K., FIELD, M., GREENING, S., GALLINGER, S., ARONSON, M., PERRIER, R., WOODS, M. O., GREEN, J. S., WALKER, N., ROSTY, C., PARRY, S. & YOUNG, J. P. 2012. Cancer Risks for Relatives of Patients With Serrated Polyposis. *Am J Gastroenterol,* 107**,** 770-778.

WITKOWSKI, L., CARROT-ZHANG, J., ALBRECHT, S., FAHIMINIYA, S., HAMEL, N., TOMIAK, E., GRYNSPAN, D., SALOUSTROS, E., NADAF, J., RIVERA, B., GILPIN, C., CASTELLSAGUÉ, E., SILVA-SMITH, R., PLOURDE, F., WU, M., SASKIN, A., ARSENEAULT, M., KARABAKHTSIAN, R. G., REILLY, E. A., UELAND, F. R., MARGIOLAKI, A., PAVLAKIS, K., CASTELLINO, S. M., LAMOVEC, J., MACKAY, H. J., ROTH, L. M., ULBRIGHT, T. M., BENDER, T. A., GEORGOULIAS, V., LONGY, M., BERCHUCK, A., TISCHKOWITZ, M., NAGEL, I., SIEBERT, R., STEWART, C. J. R., ARSENEAU, J., MCCLUGGAGE, W. G., CLARKE, B. A., RIAZALHOSSEINI, Y., HASSELBLATT, M., MAJEWSKI, J. & FOULKES, W. D. 2014. Germline and somatic SMARCA4 mutations characterize small cell carcinoma of the ovary, hypercalcemic type. *Nature Genetics,* 46**,** 438.

WOODS, M. O., WILLIAMS, P., CAREEN, A., EDWARDS, L., BARTLETT, S., MCLAUGHLIN, J. R. & YOUNGHUSBAND, H. B. 2007. A new variant database for mismatch repair genes associated with Lynch syndrome. *Hum Mutat,* 28**,** 669-73.

WORLD HEALTH ORGANISATION. 2018. *Cancer Fact Sheets: Colorectal Cancer* [Online]. International Agency for Research on Cancer. Available: https://gco.iarc.fr/today/fact-sheets-cancers [Accessed 1st October 2018].

YAMAUCHI, T., WATANABE, M., HASEGAWA, H., YAMAMOTO, S., ENDO, T., KABESHIMA, Y., YOROZUYA, K., YAMAMOTO, K. & KITAJIMA, M. 2002. Serrated adenoma developing into advanced colon cancer in 2 years. *Journal of Gastroenterology,* 37**,** 467-470.

YAN, H. H. N., LAI, J. C. W., HO, S. L., LEUNG, W. K., LAW, W. L., LEE, J. F. Y., CHAN, A. K. W., TSUI, W. Y., CHAN, A. S. Y., LEE, B. C. H., YUE, S. S. K., MAN, A. H. Y., CLEVERS, H., YUEN, S. T. & LEUNG, S. Y. 2016. RNF43 germline and somatic mutation in serrated neoplasia pathway and its association with BRAF mutation. *Gut*.

YAO, T. P., OH, S. P., FUCHS, M., ZHOU, N. D., CH'NG, L. E., NEWSOME, D., BRONSON, R. T., LI, E., LIVINGSTON, D. M. & ECKNER, R. 1998. Gene dosage-dependent embryonic development and proliferation defects in mice lacking the transcriptional integrator p300. *Cell,* 93**,** 361-72.

YOU, Y. N., XING, Y., FEIG, B. W., CHANG, G. J. & CORMIER, J. N. 2012. Young-onset colorectal cancer: is it time to pay attention? *Arch Intern Med,* 172**,** 287-9.

YOUNG, J. P., WIN, A. K., ROSTY, C., FLIGHT, I., RODER, D., YOUNG, G. P., FRANK, O., SUTHERS, G. K., HEWETT, P. J., RUSZKIEWICZ, A., HAUBEN, E., ADELSTEIN, B. A., PARRY, S., TOWNSEND, A., HARDINGHAM, J. E. & PRICE, T. J. 2015. Rising incidence of early-onset colorectal cancer in Australia over two decades: report and review. *J Gastroenterol Hepatol,* 30**,** 6-13.

YUAN, Z. M., HUANG, Y., ISHIKO, T., NAKADA, S., UTSUGISAWA, T., SHIOYA, H., UTSUGISAWA, Y., YOKOYAMA, K., WEICHSELBAUM, R., SHI, Y. & KUFE, D. 1999. Role for p300 in stabilization of p53 in the response to DNA damage. *J Biol Chem,* 274**,** 1883-6.

ZANKE, B. W., GREENWOOD, C. M., RANGREJ, J., KUSTRA, R., TENESA, A., FARRINGTON, S. M., PRENDERGAST, J., OLSCHWANG, S., CHIANG, T., CROWDY, E., FERRETTI, V., LAFLAMME, P., SUNDARARAJAN, S., ROUMY, S., OLIVIER, J. F., ROBIDOUX, F., SLADEK, R., MONTPETIT, A., CAMPBELL, P., BEZIEAU, S., O'SHEA, A. M., ZOGOPOULOS, G., COTTERCHIO, M., NEWCOMB, P., MCLAUGHLIN, J., YOUNGHUSBAND, B., GREEN, R., GREEN, J., PORTEOUS, M. E., CAMPBELL, H., BLANCHE, H., SAHBATOU, M., TUBACHER, E., BONAITI-PELLIE, C., BUECHER, B., RIBOLI, E., KURY, S., CHANOCK, S. J., POTTER, J., THOMAS, G., GALLINGER, S., HUDSON, T. J. & DUNLOP, M. G. 2007. Genome-wide association scan identifies a colorectal cancer susceptibility locus on chromosome 8q24. *Nat Genet,* 39**,** 989-94.

ZHANG, B., JIA, W. H., MATSUDA, K., KWEON, S. S., MATSUO, K., XIANG, Y. B., SHIN, A., JEE, S. H., KIM, D. H., CAI, Q., LONG, J., SHI, J., WEN, W., YANG, G., ZHANG, Y., LI, C., LI, B., GUO, Y., REN, Z., JI, B. T., PAN, Z. Z., TAKAHASHI, A., SHIN, M. H., MATSUDA, F., GAO, Y. T., OH, J. H., KIM, S., AHN, Y. O., CHAN, A. T., CHANG-CLAUDE, J., SLATTERY, M. L., GRUBER, S. B., SCHUMACHER, F. R., STENZEL, S. L., CASEY, G., KIM, H. R., JEONG, J. Y., PARK, J. W., LI, H. L., HOSONO, S., CHO, S. H., KUBO, M., SHU, X. O., ZENG, Y. X. & ZHENG, W. 2014. Large-scale genetic study in East Asians identifies six new loci associated with colorectal cancer risk. *Nat Genet,* 46**,** 533-42.

ZHANG, J., AHMAD, S. & MAO, Y. 2007. BubR1 and APC/EB1 cooperate to maintain metaphase chromosome alignment. *The Journal of Cell Biology,* 178**,** 773-784.

ZHANG, J., SUN, X., QIAN, Y., LADUCA, J. P. & MAQUAT, L. E. 1998a. At Least One Intron Is Required for the Nonsense-Mediated Decay of Triosephosphate Isomerase mRNA: a Possible Link between Nuclear Splicing and Cytoplasmic Translation. *Molecular and Cellular Biology,* 18**,** 5272-5283.

ZHANG, J., SUN, X., QIAN, Y. & MAQUAT, L. E. 1998b. Intron function in the nonsense-mediated decay of beta-globin mRNA: indications that pre-mRNA splicing in the nucleus can influence mRNA translation in the cytoplasm. *RNA,* 4**,** 801-815.

ZHANG, J. X., FU, L., DE VOER, R. M., HAHN, M. M., JIN, P., LV, C. X., VERWIEL, E. T., LIGTENBERG, M. J., HOOGERBRUGGE, N., KUIPER, R. P., SHENG, J. Q. & GEURTS VAN KESSEL, A. 2015. Candidate colorectal cancer predisposing gene variants in Chinese early-onset and familial cases. *World J Gastroenterol,* 21**,** 4136-49.

ZIMPRICH, A., BISKUP, S., LEITNER, P., LICHTNER, P., FARRER, M., LINCOLN, S., KACHERGUS, J., HULIHAN, M., UITTI, R. J., CALNE, D. B., STOESSL, A. J., PFEIFFER, R. F., PATENGE, N., CARBAJAL, I. C., VIEREGGE, P., ASMUS, F., MÜLLER-MYHSOK, B., DICKSON, D. W., MEITINGER, T., STROM, T. M., WSZOLEK, Z. K. & GASSER, T. 2004. Mutations in LRRK2 Cause Autosomal-Dominant Parkinsonism with Pleomorphic Pathology. *Neuron,* 44**,** 601-607.