Copy number variants and their role in hereditary breast cancer and hereditary colorectal cancers

Amy Louise Masson

GradDipForStForSc, BBioMedSci (Hons), BSc

Doctor of Philosophy, Medical Genetics

The University of Newcastle, Australia

September 2015

Declarations

Statement of originality

The thesis contains no material which has been accepted for the award of any other degree or diploma in any university or other tertiary institution and, to the best of my knowledge and belief, contains no material previously published or written by another person, except where due reference has been made in the text. I give consent to the final version of my thesis being made available worldwide when deposited in the University's Digital Repository, subject to the provisions of the Copyright Act 1968.

Statement of collaboration

I hereby certify that the work embodied in this thesis has been done in collaboration with other researchers. I have included as part of the thesis a statement clearly outlining the extent of collaboration, with whom and under what auspices.

Statement of authorship

I hereby certify that the work embodied in this thesis contains a published paper/s/scholarly work of which I am a joint author. I have included as part of the thesis a written statement, endorsed by my supervisor, attesting to my contribution to the joint publication/s/scholarly work.

Thesis by publication

I hereby certify that this thesis is in the form of a series of published papers of which I am a joint author. I have included as part of the thesis a written statement from each co-author, endorsed by the Faculty Assistant Dean (Research Training), attesting to contribution to the joint publications.

Amy Louise Masson Date: 01/09/2015

Acknowledgements

There are many people whom without their assistance this work would not have been successfully completed and to which I owe many thanks.

I would firstly like to thank my principle supervisors Professor Rodney Scott and Doctor Bente Talseth-Palmer for providing me with the opportunity to undertake this candidature and for sharing with me their knowledge, experience and expertise to ensure it was successfully completed.

I am deeply grateful for all the support and advice from everyone at Information Based Medicine at the Hunter Medical Research Institute and the staff at the Molecular Genetics Department of the Hunter Area Pathology Service. To all, without your ongoing support and encouragement it would have made this work all the much harder, and certainly less enjoyable. I would especially like to thank Desma Grice, Trish Collinson, Melissa Tooney, Tiffany-Jane Evans, David Mossman and Michelle Wong-Brown for their generosity in the time put towards my project including reading my many manuscript and thesis drafts, answering lots of pesky question as well as providing me with valuable feedback that kept me pointed in the right direction.

I would like to thank all my family and friends, with an especially big thank you to my parents who encouraged me to go through university and unknowingly helped me take my first of so many steps that have lead me to where I am today. Thank you for giving me the wisdom and courage to make those choices, the unfailing support you have provided every single day of my life and encouraging me to always do my best. Thank you for always being there and I hope I have made you proud.

Lastly I would like to thank my husband, Stuart Masson for which your patience, love and support has meant so much. Without you by my side and encouraging me to be brave and take the leap to do this work, I would be somewhere very different now.

Thank you.

I dedicate this work to my daughter Ainslie Evelyn Masson
who reminds me every day what determination is.

List of publications included as part of this thesis

Published original research articles

Amy L. Masson, Bente A. Talseth-Palmer, Tiffany-Jane Evans, Desma M. Grice, Garry N. Hannan and Rodney J. Scott (2014) Expanding the genetic basis of copy number variation in familial breast cancer, *Hereditary Cancer in Clinical Practice*, 12:15.

Amy L. Masson, Bente A. Talseth-Palmer, Tiffany-Jane Evans, Desma M. Grice, Konsta Duesing, Garry N. Hannan and Rodney J. Scott (2013) Copy Number Variation in Hereditary Non-Polyposis Colorectal Cancer, *Genes*, *4*, 536-555.

Submitted manuscripts

Amy L. Masson, Bente A. Talseth-Palmer and Rodney J. Scott (2015) Interrogation of genes disrupted by copy number variants in familial breast cancer using a bioinformatics approach, *International Journal of Cancer Research and Diagnosis*.

Amy L. Masson, Bente A. Talseth-Palmer, Tiffany-Jane Evans, Patrick McElduff, Allan D. Spigelman, Garry N. Hannan and Rodney J. Scott (2015) Copy number variants associated with 18p11.32, *DCC* and the promoter 1B region of *APC* in familial adenomatous polyposis, *Gene*.

Amy L. Masson, Bente A. Talseth-Palmer, Tiffany-Jane Evans, Trish Collinson, Michelle Wong-Brown, Melissa A. Tooney, Garry N. Hannan and Rodney J. Scott (2015) Intronic variants resulting in aberrant mRNA species are rare in Hereditary Non-Polyposis Colorectal Cancer, *Hereditary Cancer in Clinical Practice.*

Copyright statement

I warrant that I have obtained, where necessary, permission from the copyright owners to use any third party copyright material produced in the thesis (e.g. questionnaires, artwork, unpublished letters) or to use my own published work (e.g. journal articles) in which the copyright is held by another party (e.g. publisher, co-author).

Amy Louise Masson Date: 01/09/2015

List of additional material relevant to the thesis but not forming part of it

Publications

Bente A. Talseth-Palmer, Elizabeth G. Holliday, Tiffany-Jane Evans, Mark McEvoy, John Attia, Desma M. Grice, Amy L. Masson, Cliff Meldrum, Allan Spigelman and Rodney J. Scott (2013) Continuing difficulties in interpreting CNV data: lessons from a genome-wide CNV association study of Australian HNPCC/Lynch syndrome patients, *BMC Medical Genomics, 6*(*10*), 1-13.

Conference presentations

Amy L. Masson, Bente A. Talseth-Palmer, Tiffany-Jane Evans, Desma M. Grice, Konsta Duesing, Garry N. Hannan and Rodney J. Scott (2013) Expanding the genetic basis of hereditary non polyposis colorectal cancer. *InSight Meeting*, Cairns, Australia.

Amy L. Masson, Bente A. Talseth-Palmer, Desma M. Grice, Konsta Duesing, Garry N. Hannan and Rodney J. Scott (2012) A comprehensive catalogue of copy number variants in hereditary colorectal cancers. *Biomarker Discovery Conference*, Shoal Bay, Australia.**

Amy L. Masson, Bente A. Talseth-Palmer, Desma M. Grice, Garry N. Hannan and Rodney J. Scott (2012) Copy number variation in hereditary colorectal cancer. *American Society for Human Genetics*, San Francisco, United States of America.

Amy L. Masson, Bente A. Talseth-Palmer, Desma M. Grice, Garry N. Hannan and Rodney J. Scott (2012) Copy number variation and its role in hereditary non-polyposis colorectal cancer. *Australian Society for Medical Research NSW scientific meeting*, Sydney, Australia.

Amy L. Masson, Bente A. Talseth-Palmer, Desma M. Grice, Garry N. Hannan and Rodney J. Scott (2012) Elucidating the genetic predisposition to familial adenomatous polyposis. *Human Genome Meeting*, Sydney, Australia.

Amy L. Masson, Bente A. Talseth-Palmer, Desma M. Grice, Garry N. Hannan and Rodney J. Scott (2011) Elucidating the genetic predisposition to colorectal cancer. *Hunter Medical Research Institute Cancer Research Program Symposium*, Newcastle, Australia.*

Amy L. Masson, Bente A. Talseth-Palmer, Desma M. Grice, Garry N. Hannan and Rodney J. Scott (2011) Elucidating the genetic predisposition to colorectal cancer. *Australian Society for Medical Research NSW scientific meeting*, Sydney, Australia.

**invited presentation*

***oral presentation*

Table of Contents

Table of Figures

Table of Tables

ABSTRACT

Hereditary breast cancer and hereditary colorectal cancers are associated with an earlier age of diagnosis and a higher frequency of disease among family members. In recent decades cancer susceptibility genes have been associated with hereditary forms of breast cancer and colorectal cancers however these genes only account for a minority of families seeking diagnostic testing.

Genetic variation explains a significant proportion to susceptibility of disease. Copy number variants (CNVs) are a form of structural genetic variation yet to be fully explored for their contribution to hereditary breast cancer or hereditary colorectal cancers. CNV analysis can be used to identify new genes and loci which may be associated with disease risk.

The Affymetrix Cytogenetic Whole Genome 2.7M (Cyto2.7M) array was used to detect regions of genomic gain and loss in a cohort of 350 samples (encompassing 129 *BRCA1*/*BRCA2* mutation negative hereditary breast cancer patients, 56 Familial adenomatous polyposis (FAP) *APC* mutation negative and 125 Hereditary non-polyposis colorectal cancer (HNPCC) mismatch repair (MMR) mutation negative colorectal cancer patients and each were compared to 40 healthy control genomes).

CNV analysis revealed the presence of 614 genes unique to the combined patient cohort which represent candidates for involvement in hereditary breast cancer and hereditary colorectal cancers. Several CNVs were found that were associated with previously reported cancer susceptibility genes. These included CNVs associated with *APC*, *DCC*, *MLH1* and *CTNNB1* in four polyposis patients and *RPA3*, *NBN* (*NBS1*), *MRE11A* and *CYP19A1* in five breast cancer patients and suggests their role in disease development in the affected individuals. Of special interest was the identification of *WWOX* and *FHIT* rearrangements in three breast cancer patients, and a recurrent deletion that was observed on chromosome 18 at position 18p11.32 in 9% of the polyposis patients screened. These variants could further account for disease in the affected patients. Bioinformatic analysis of the uniquely identified gene sets provided further insight into the roles of these genes in disease.

This thesis provides evidence supporting the hypothesis that CNVs are likely contributors to disease development in a small but significant proportion of hereditary breast cancer and hereditary colorectal cancer patients.

CHAPTER 1: BACKGROUND

Introduction

Genetic variation explains a significant proportion of susceptibility to common disease[1-5]. The identification of specific genetic variants associated with disease is therefore a priority in delineating the pathological processes which underlie complex diseases such as cancer[6].

Genetic variation ranges from single base pair (bp) changes through to large-scale structural alterations[1]. All genetic changes have the potential to alter gene expression which can lead to a significant change in a cells malignant potential.

Copy number variation (CNV) describes a form of structural variation that encompasses genomic events termed duplications and deletions (or the gain or loss of genomic material)[7] and was first described in 1936 by Bridges *et al.* who discovered the 'bar' gene duplication in *Drosophila*[8]. It wasn't until ~70 years later however, when technologies emerged allowing for the first time high fidelity analysis of the whole genome, did significant research into the presence and role of CNVs in the genome and their relevance disease[9-12].

It is through understanding the mechanisms involved in cancer development and progression that disease incidence may be reduced and furthermore the knowledge to manage, treat and potentially cure cancer may arise. In 1952, Nordling[13] first put forward the theory that successive DNA mutations are the cause of cancer and that the frequency of cancer should increase in direct proportion to age (as mutated cells increase in number, so does the probability of cancer developing). This relationship can be observed in diseases such as skin cancer (especially squamous cell carcinomas and basal cell carcinomas) where ultra-violet (UV) light exposure correlates with disease incidence; and in lifestyle choices like cigarette smoking where carcinogen exposure is correlated with cancers of the lung, oral cavity, larynx and oesophagus[14,15]. See figure 1 illustrating the interconnectivity of factors involved in cancer development.

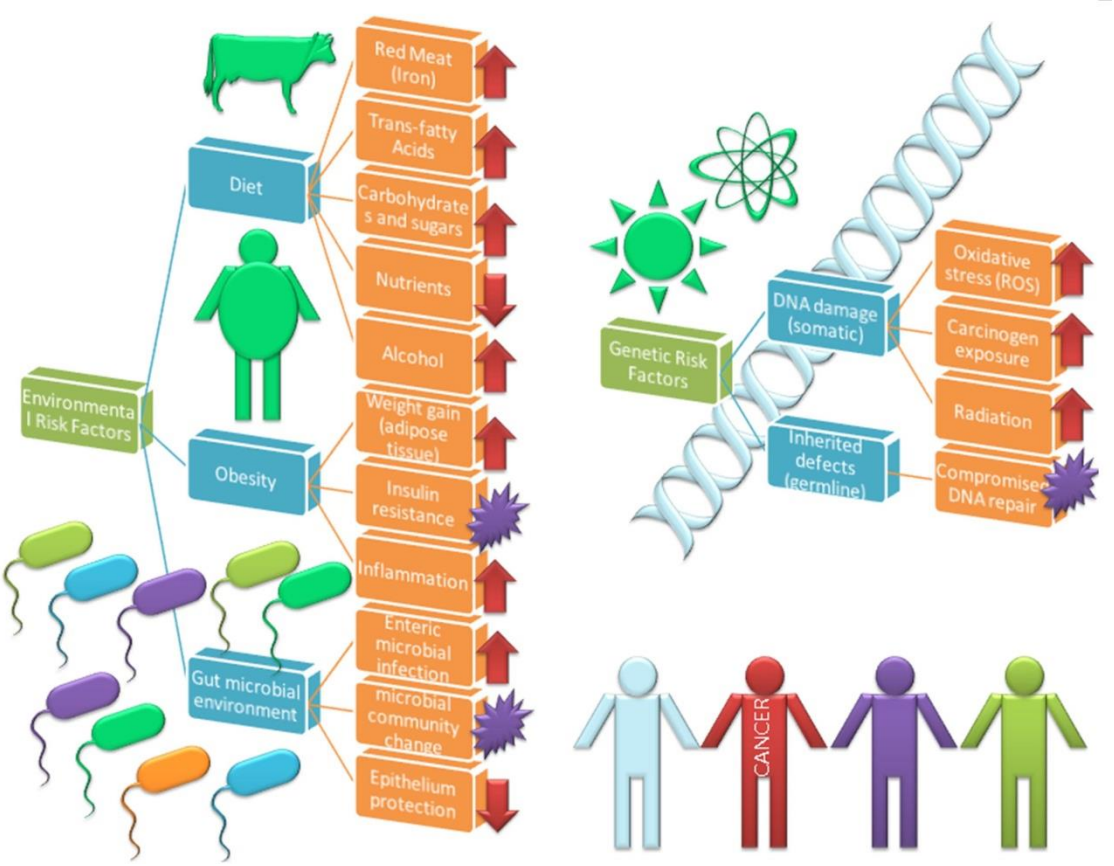


Figure 1 Interconnectivity of environmental and genetic factors in colorectal cancer development

While DNA damage is a common and sometimes unavoidable occurrence in individuals throughout life, mechanisms have evolved that have rendered cells with the capacity to repair DNA damage and protect against disease development. DNA repair is orchestrated by any of six main repair pathways: direct reversal (DR), base excision repair (BER), mismatch repair (MMR), nucleotide excision repair (NER), non-homologous end-joining (NHEJ) and homologous recombination (HR) depending on what type of lesion is created[16]. Each of these repair pathways targets different types (or in some cases, a combination) of DNA damage, for example: the MMR pathway repairs DNA mismatches, while the NHEJ and HR pathways repair double-strand breaks (DSBs)[17,18]. These pathways are essential in reducing global DNA damage; however when repair capacity is compromised, genomic integrity and cell fidelity is reduced which can result in genomic instability, apoptosis, cell senescence, or sometimes malignancy.

Several malignancies have been unequivocally associated with compromised DNA repair, of which hereditary breast cancer and hereditary colorectal cancers are the primary focus in this body of work.

Colorectal cancer is considered a complex disease and one of the most common cancers affecting affluent societies. It represents the third most commonly diagnosed cancer in males and the second most common in females[19]. The highest incidences of disease are observed in Scotland, Australia, New Zealand, Europe and North America, with an estimated incidence of 10% of all new cancers diagnosed and is associated with ~8% of all cancer deaths[19].

Somewhere between 70% and 80% of all colorectal cancers occur without a family history and are considered to be a result of both genetic and environmental factors, typically affecting individuals in their sixth and seventh decades of life[20-23]. In the remaining 20-30% of cases, colorectal cancer appears to be familial with a percentage of these having an distinct inherited pattern of disease transmission[21,24]. Less than 6% of all colorectal cancers are due to highly penetrant mutations occurring in a set of defined genes[22,25].

Breast cancer represents the most commonly diagnosed female malignancy[19] . Since early in the new millennia, breast cancer susceptibility has been described as being heightened in individuals who conform to hereditary colorectal cancer syndromes, such as hereditary non-polyposis colorectal cancer (HNPCC)[26-28]. It is considered that the colon or rectal tumours of these breast cancer patients display similar genetic features

such as microsatellite instability (MSI) which indicates they may belong to the same disease entity (reviewed in[28]).

Alone, breast cancer is estimated to account for approximately 25% of new cases and 15% of cancer deaths, respectively, with greater than 60% of breast cancer deaths specifically transpiring in developing countries despite many of these countries having low to intermediate levels of disease incidence[19]. Industrialized countries exhibit a high incidence of breast cancer but lower death rates which are widely accepted to be attributed to the use of hormone replacement therapies (HRT) and the abundance of early detection programs[19]. In fact HRT use is reported to be associated with a greater than 20% increase in breast cancer risk and this risk may further vary depending on the patient's race, body mass index (BMI) and breast density[29].

Hereditary breast cancer and hereditary colorectal cancers, are associated with mutations arising in known cancer susceptibility genes, however for the majority of patients seeking genetic testing for their condition, disease cannot be explained by any obvious coding changes in the relevant genes tested. This suggests that either other genes yet to be identified or other mechanisms of gene inactivation may be responsible for disease in these patients. The role of CNVs in hereditary breast cancer and hereditary colorectal cancers is yet to be fully elucidated, suggesting CNVs may account for disease in a proportion of patients.

Colorectal Cancer

The colon represents one of the major components of the vertebrate digestive system (see figure 2), and is comprised of four distinct regions: the ascending, transverse, descending and sigmoid colon. The large intestine also features the appendix, and terminates to the rectum and anus.

The epithelium of the colon is protected by the colonic mucosa which is the primary physical barrier to the underlying cells (for a full review see[30]). Development of the mucosal barrier is driven by the presence of endemic communities of bacteria and microorganisms which reside in a healthy gastrointestinal (GI) tract[31-33]. Evidence in the literature suggests constituents of diets are responsible for altering the composition of microbiota residing in the GI tract and this may be related to gut health and disease risk[34-36]. Ingestion of apple pectin, for example, has been found to increase the compound butyrate and b-glucuronidase producing the growth of *Clostridiales*, whilst also resulting in a decrease of several species of *Bacteroidetes*[35]. Changes in the composition of species residing in the GI tract have been observed in intestinal diseases and are suggested to play a role in disease pathogenesis[37]. Investigations have identified a similar association between diet and gut microbial communities and the risk of colorectal cancer development, with one potential mechanism of this being the activation of anti-proliferative and anti-inflammatory molecules[38-41].

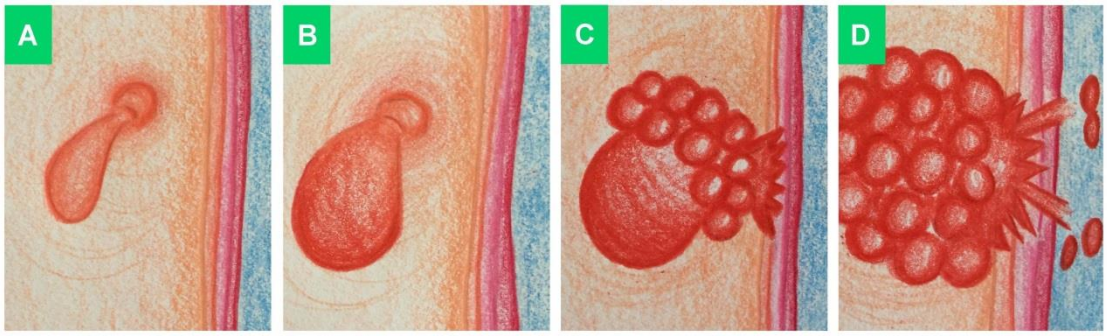
The large colon is potentially exposed to the highest concentrations of exogenous compounds compared to other regions in the gastrointestinal tract. This is due to it being the site for nutrient absorption and fluid salvation[42,43], which results in the concentration of potentially harmful constituents such as elementary iron (e.g. from meat) and trans-fats acids[44-53]. Iron consumption has been reported to increase the risk of colorectal cancer development in Hereditary Haemochromatosis patients; in addition to amplifying WNT signalling, which in the presence of a mutant *APC* gene can exacerbate tumorigenesis in FAP patients[50-52]. Reactive oxygen species (ROS) are generated during iron digestion and this is thought to be the mechanism by which elevated iron may contribute to DNA damage and malignancy (for a full review see[53]).

Total fat intake has also been attributed to the development of cancer including colorectal cancer, although measures of total dietary fat intake have shown no association[54-60]. Diets which contain high levels of trans-fatty acids have been attributed to an increased BMI as well as an increased risk of developing colorectal cancer[45-49]. Further studies have identified this relationship to be strongly correlated

with the development of sigmoid colon, rectosigmoid and rectum carcinomas specifically[60]. The mechanism by which trans-fatty acids may contribute to colorectal cancer development is also suggested to be through the generation of ROS which are produced as a bi-product during the digestion of trans-fatty acids[61].

Gut microbial community composition and diet, along with the long transit times experiences through the colon, maximizes exposure to carcinogens and it is a commonly held view that this exacerbates disease risk. Direct contact between exogenous agents in the colonic lumen and the epithelium may still occur despite the mucosal barrier and it is this contact that is believed to be responsible for inflicting cellular damage, including DNA damage that ultimately results in malignancy[42,43].

Susceptibility to developing cancer is influenced by the genetic background of an individual. Hereditary colorectal cancer syndromes are classified according to their respective clinical phenotypes and the genes involved in predisposing affected individuals to disease development (for a full review[52] and see table 1 for a summary). Several hereditary cancer syndromes have also been described that are unequivocally associated with reduced or absent DNA repair, most of which are relatively rare however included among these cancers is the most common predisposition to colorectal cancer, HNPCC[62].



(A) Cells in the epithelium of the colon constantly divide, some of which become hyperplastic and continue to grow excessively; (B) hyperplasia leads to the development of an adenoma which accumulates genetic changes which can turn it into a carcinoma; (C) the carcinoma eventually breaks through into the muscle wall surrounding the colon, gathering more genetic changes as it grows; and (D) finally the cancerous cells break through the muscular wall surrounding the colon allowing the cancer to spread to other parts of the body to develop distant metastasis.

Figure 2 Illustration of colorectal cancer development

Table 1 Types of hereditary colorectal cancers.

Disease	Genes	References
Non-polyposis syndromes		
Hereditary non-polyposis colorectal cancer (HNPCC)/Lynch syndrome (LS) Muir Torre Syndrome	*MSH2* *MSH6* *MLH1* *PMS2*	63-74
Turcot's syndrome	*MSH2* *MSH6* *MLH1* *PMS2* *APC*	75
Polyposis Syndromes		
Adenomatous polyposis		
Familial adenomatous polyposis (FAP)	*APC*	76-79
MUTYH associated polyposis (MAP)	*MUTYH*	80-83
Hamartomatous polyposis		
Juvenile polyposis (JP)	*PTEN* *SMAD4* *BMPR1A*	84,85
Bannayan-Riley-Ruvalcaba syndrome and Cowden disease	*PTEN*	86
Peutz-Jeghers syndrome (PJS)	*STK11*	87,88
Hyperplastic polyposis		
Hyperplastic polyposis	*unknown*	52
Other polyposis		
Hereditary mixed polyposis syndrome (IMPS), Inflammatory bowel disease/Crohn's disease	*various*	52

Familial adenomatous polyposis (FAP)

FAP is an autosomal dominant inherited colorectal cancer syndrome affecting approximately 1 in 10,000-12,000 individuals and accounts for less than 0.5% of all colorectal cancers[21,52,89,90]. It is the second most common form of hereditary colorectal cancer and it has a severe phenotype[91]. Defects in the tumour suppressor gene *APC* have been identified to be associated with disease in the majority of FAP and attenuated FAP patients and for which over 1400 pathogenic mutations have been described[92]. In 2002, mutations in the BER gene *MUTYH* were also found to account for an autosomal recessive form of FAP, specifically referred to as *MUTYH* associated polyposis (MAP)[93-95] (see[96] for a recent review on MAP). In the classical form, FAP is characterized by early development of hundreds to thousands of adenomas in the colon and rectum, which commonly develop during early childhood and adolescence with one or more adenomas becoming malignant a decade later, resulting in colorectal cancer[90]. The average age of colorectal cancer diagnosis in FAP is 35-40 years and there is almost 100% disease penetrance[21,76,89,90].

Extra-colonic manifestations such as duodenal adenomas, polyps of the fundic gland, desmoids tumours, osteomas and lipomas are also associated with FAP[76,90]. In particular, desmoids tumours are considered the second major cause of morbidity and mortality in FAP patients, accounting for approximately 10% of patients[97,98]. Gene expression profiling of desmoids tumours suggest similarities exist in the mechanisms involved in tumour formation that are also important in the development of colorectal tumours in classical FAP patients[97,98]. The less severe forms of FAP (MAP and attenuated FAP) characteristically involve the development of fewer adenomas and are associated with a later age of disease onset[99].

Molecular testing of FAP

In 80-90% of FAP pateints, germline sequence variants are identified in the *APC* gene which results in a truncated non-functional gene protein[100]. Most commonly these occur either in the 5' end of the *APC* gene, the Mutation Cluster Region (MCR), or at codons 1309 and 1450 which alone accounts for 35% of all variants identified[76,101-103]. Approximately 2% of all aberrations identified are large gene deletions, including those which extend from the promoter into the coding region of the gene[104,105]. *APC* is reported to have two promoter regions (1A and 1B) with the latter of these proposed to have only a minor role in regulating *APC* gene expression[106-111]. Recent evidence has however, suggested that disease variants in either promoter can inactivate *APC* and

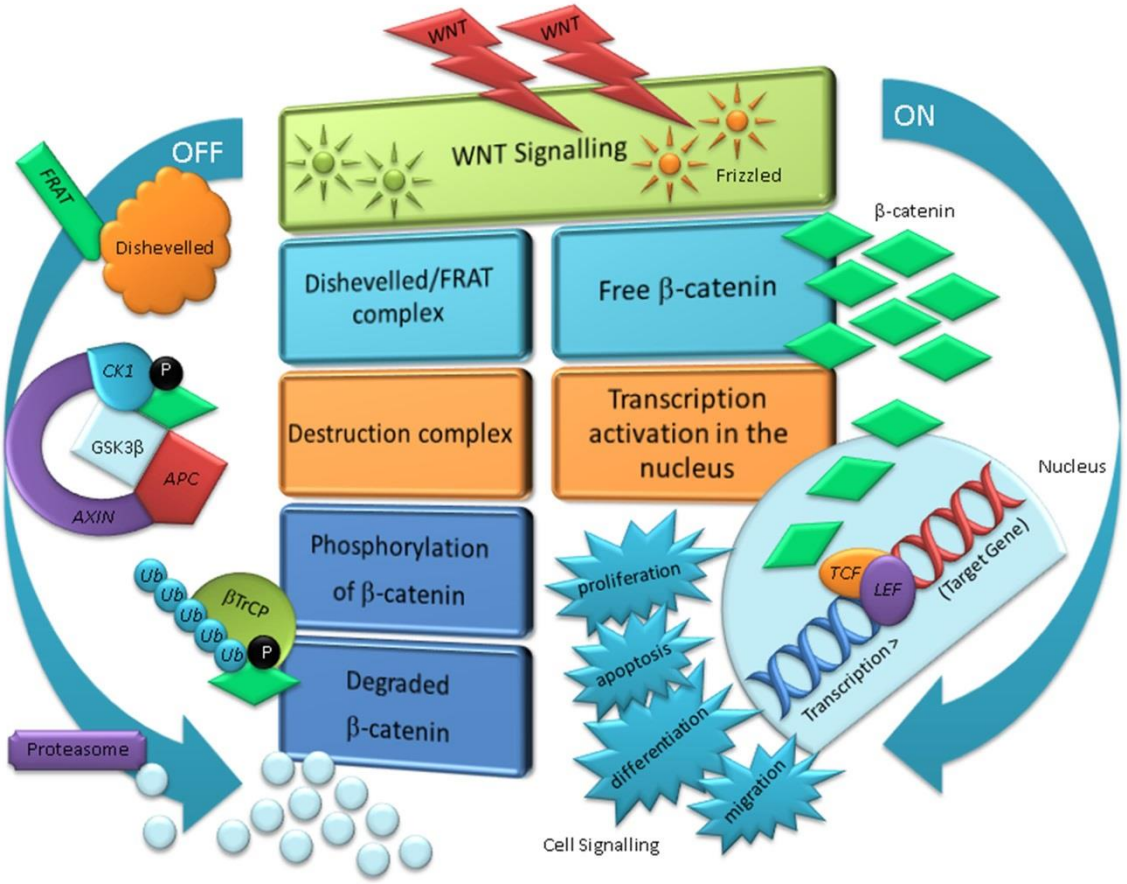
can result in disease[106,112]. The *APC* gene also expresses multiple gene transcripts, for which alternatively or aberrantly spliced examples have been associated with FAP[113]. Furthermore, aberrations in codon 157 of *APC* have been specifically associated with the development of attenuated FAP[76,101] .

Mutations in *MUTYH* are reported to account for approximately 10% of patients with classical FAP and up to 30% of attenuated FAP patients[94]. The InSiGHT database contains over 300 known pathogenic variants in *MUTYH*, of which the two missense mutations c.536A>G (Y179C) and c.1187G>A (G396A) are estimated to account for 50% and 82%, respectfully, of the variants identified[114,115]. Observations in the literature suggest these variants have arisen from founder affects that occurred prior 5,000 BC in Europe[115]. It should also be noted that low levels of somatic mosaicism is known to be an alternative cause of FAP in a minority of patients[116-120].

The WNT Signalling Pathway

The WNT signalling pathway, specifically the canonical sub-pathway, is well characterized in its involvement with cancer (see figure 3)[89,90121,122]. The WNT signalling pathway is involved in the regulation of nuclear β-catenin levels responsible for the constitutive activation of a number of transcription factors and the regulation of several gene expression networks (cell proliferation, cell polarity, cell fate determination during embryonic development and tissue homeostasis[123,124]) that are directly involved in FAP development[95,96]. As *APC* is an antagonist in the WNT signalling pathway[125], disruption of this gene can increase pathway activation, resulting in unregulated cell signalling leading to aberrant cellular differentiation, apoptosis, migration and proliferation[125-127].

Approximately 80% of FAP families are found to have a genetic variant residing in *APC* and up to 10% have mutations in *MUTYH* leaving the remaining 10% with an unresolved genetic cause of disease. This therefore suggests that other genetic variants associated with *APC* or *MUTYH* are yet to be identified which may give rise to FAP, or alternatively, that other genes (such as other genes residing in the canonical WNT signalling pathway) may also be involved in the aetiology of this disease.



(ON) WNT genes signal to trans-membrane frizzled receptors which activate the dishevelled complex of proteins and result in the phosphorylation of GSK3β. β-catenin is free to translocate to the nucleus. Here it associates with transcriptional machinery resulting in the activation of several downstream genes.

(OFF) In the presence of no WNT signalling, GSK3β forms part of the β-catenin destruction complex, responsible for the phosphorylation of β-catenin and its subsequent degradation by the cell proteasome. This inhibits the translocation of β-catenin to the nucleus and prevents the activation of transcription targets (adapted from[126,127]).

Figure 3 Basics of the WNT signalling pathway.

Hereditary non-polyposis colorectal cancer (HNPCC)

The most common form of hereditary colorectal cancer is HNPCC, which describes a familial clustering of epithelial malignancies, most often colorectal cancer followed by endometrial cancer (in women)[128,129]. HNPCC was originally described by Warthin in 1913 and later defined by Lynch in 1966[130]. By definition, HNPCC describes families that conform to the Amsterdam Criteria/Bethesda guidelines. The Amsterdam criteria (1990) was initially developed to assist in the identification of genes associated with the predisposition, though has been adopted today to assist in determining a clinical diagnosis[131]. The Amsterdam criteria was however deemed insufficient as it was considered too stringent to identify HNPCC families with extra colonic cancers (especially endometrial cancer)[132]. As a result, the Bethesda guidelines and the Amsterdam Criteria II were devised to accommodate families with extra colonic phenotypes (see table 2).

Lynch syndrome (LS) is a subset of HNPCC describing an autosomal dominant inherited disorder associated with defects in DNA MMR genes[63-69,73]. Prior to the identification of the genetic basis of LS, all such families were defined as HNPCC. HNPCC/LS are associated with an increased risk of colorectal cancer at earlier than expected ages and most recent disease penetrance estimations suggest by the age of 70 years, 53% of men and 33% women will develop colorectal cancer, and 44% of women will develop endometrial cancer[128]. In comparison, lifetime risk of colorectal cancer in the general population is only 4%[133,134]. Individuals diagnosed with HNPCC/LS are also at increased risk of developing other epithelial cancers including small bowel, gastric, ovarian, hepatobiliary tract, urologic tract and brain tumours[22,62,134-136].

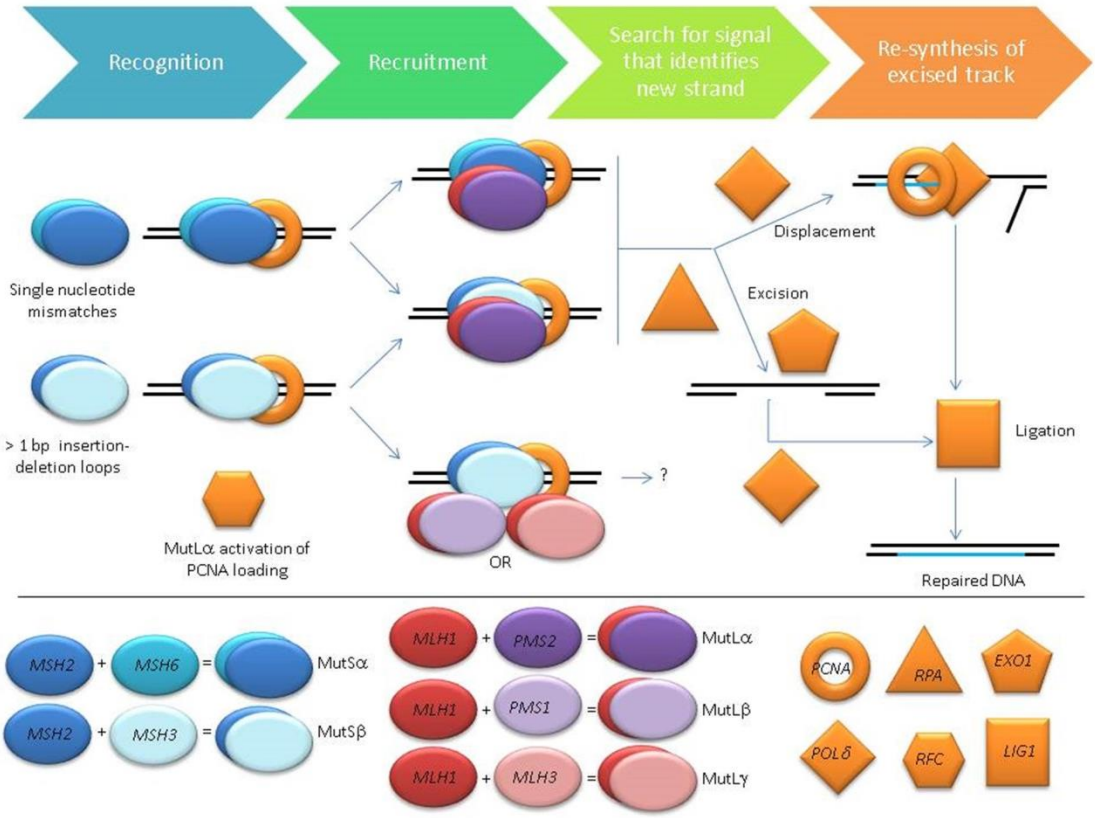
Table 2 Summary of the Amsterdam criteria's and Bethesda guidelines.

Amsterdam Criteria[131] (all criteria must be met)	
1	One member diagnosed with colorectal cancer before age 50 years
2	Two affected generations
3	Three affected relatives, one of them a first-degree relative of the other two
4	FAP should be excluded
5	Tumours should be verified by pathologic examination
Amsterdam Criteria II[137] (all criteria must be met)	
1	There should be at least three relatives with HNPCC associated cancer (colorectal cancer or cancer of the endometrium, small bowel, urethra, or renal pelvis)
2	Should be a first-degree relative of the other two
3	At least two successive generations should be affected
4	At least one should be diagnosed by the age of 50 years
5	FAP should be excluded
6	Tumours should be verified by pathologic examination
Bethesda guidelines[138,139] (meeting features listed under any of the numbered criteria is sufficient)	
1	Individuals with cancer in families that meet the Amsterdam criteria
2	Individuals with two HNPCC-related cancers, including synchronous and metachronous colorectal cancers or associated extra colonic cancers (Note: endometrial, ovarian, gastric, hepatobiliary, or small bowel cancer or transitional cell carcinoma of the renal pelvis or urethra)
3	Individuals with colorectal cancer and a first degree relative with colorectal cancer and/or HNPCC-related extra-colonic cancer and/or a colorectal cancer adenoma; one of the cancers diagnosed at age younger than 45 years, and the adenoma diagnosed at an age younger than 40 years
4	Individuals with colorectal cancer or endometrial cancer diagnosed at age younger than 45 years
5	Individuals with colorectal cancer with an undifferentiated pattern (solid/cribriform) on histopathology diagnosed at age younger than 45 years (Note: solid/cribriform defined as poorly differentiated or undifferentiated carcinoma comprised of irregular, solid sheets of large eosinophilic cells and containing small gland-like spaces)
6	Individuals with signet-ring cell type colorectal cancer diagnosed at age younger than 45 years (Note: composed of >50% signet-ring cells. A National Cancer Institute Workshop on Hereditary Non-polyposis Colorectal Cancer Syndrome: meeting highlights and Bethesda guidelines
7	Individuals with adenomas diagnosed at age younger than 40 years

The mismatch repair (MMR) pathway

Between 1993 and 1995 mutations in four DNA MMR genes (*MLH1*, *MSH2*, *MSH6* and *PMS2*) were linked to LS[63-69]. Other genes in the MMR pathway such as *MLH3*, *EXO1* and *PMS1* have also been suggested to have a role in LS, however controversial results have been reported and therefore these genes are not systemically tested in a clinical setting[140-143]. More recently, reports identify the loss of the polyadenylation signal associated with *EPCAM* to result in the transcriptional silencing of *MSH2* and this represents another mechanism in LS development[70-72,144].

MMR genes are involved in DNA repair and are responsible for a variety of genetic stabilisation functions, including error correction during DNA replication, overseeing events in genetic recombination, checkpoint and apoptotic responses[145,146]. The primary function of the MMR pathway is to eliminate base-base mismatches and insertion-deletion loops which arise as a consequence of DNA polymerase slippage during DNA replication[147]. Four stages have been suggested for MMR: (1) the recognition of mismatches; (2) the recruitment of the MMR participatory elements; (3) the search for the signal that identifies the newly synthesized DNA strand; and (4) the re-synthesis of the excised tract[148]. Across these stages, 22 genes (*MSH2*, *MSH3*, *MSH6*, *MLH1*, *MLH3*, *PMS1*, *PMS2*, *EXO1*, *RFC1*, *RFC2*, *RFC3*, *RFC4*, *RFC5*, *PCNA*, *RPA1*, *RPA2*, *RPA3*, *POLD1*, *POLD2*, *POLD3*, *POLD4* and *LIG1*) are known to participate in eukaryote MMR, each contributing to the overall repair capacity of the cell (see figure 4)[149-175]. A cell harbouring a compromised MMR gene will have a reduced MMR capacity that may result in mutations accumulating in one of several genes necessary for malignant transformation[62,148]. When no MMR is active, replication errors accumulate in (usually) dinucleotide DNA repeat sequences, resulting in MSI, the hallmark mutator phenotype of HNPCC[176,177].



MutS homologues recognize DNA mismatch errors, recruiting MutL homologues to form heterodimers specific to the type of mismatch to be repaired. These heterodimers are held on to the DNA by the PCNA clamp, which is loaded onto and off the DNA by the RFC complex of proteins. EXO1 excises the error nucleotides while the RPA complex of proteins unwinds and stabilizes the DNA throughout this process. The polymeraseδ complex generates the new (error-free) DNA strand, which is then joined back together by LIG1 (diagram adapted from[17]).

Figure 4 Flow diagram of the MMR process of DNA repair.

Molecular testing of HNPCC

The clinically significant genes in LS are *MLH1*, *MSH2*, *PMS2* and *MSH6* as they have been demonstrated to directly result in disease development[134,143,147]. When the Amsterdam Criteria is met, an aberration in any of these four genes is likely to be identified in a high proportion of suspected HNPCC cases (depending on the methodology used)[178-181]. In comparison, only 16-30% of individuals with a clinical diagnosis of HNPCC is likely to have a MMR mutation if the Bethesda guidelines are met[179,181]. At present, mutations in these four genes account for the majority of detectible germline mutations. Approximately 50% of germline mutations are detected in *MLH1*, 40% in *MSH2*, 5-10% in *MSH6* and a few families have *PMS2* mutations[143,147,182]. Variants in *EPCAM* are estimated to account for up to 3% of LS families[71,144]. Despite our understanding of HNPCC/LS disease pathology, anywhere from 30-50% of clinically tested patients will fail to have an identified germline mutations in any of the four MMR genes tested[128,129,183]. This suggests that other genetic variants yet to be identified are associated with these genes and may give rise to LS, or alternatively, that other genes not routinely screened for are involved in the aetiology of this disease, for example, other genes in the MMR pathway.

Breast Cancer

The breast represents the female reproductive organ which has the primary purpose to produce and store colostrum and later milk for mammalian offspring via a process called lactation[184]. The breast is comprised of three distinct regions being the lobules which produce and store the colostrum/milk, the ducts which carry it to the nipple as well as the fatty (adipose) connective tissue that surrounds these structures (see figure 5)[184].

The breast typically undergoes three major periods of structural change, being the embryonic, pubertal and reproductive stages which are associated with breast growth (cell replication) and reduction (cell death or apoptosis)(for a full review on mammary gland development see[184]). During puberty there is an increase in both oestrogen and progesterone which triggers the growth of the breast components to form functioning mammary glands, while during the child bearing years monthly fluctuations in firstly oestrogen (increases in the first part of the menstrual cycle) triggers the growth in the milk ducts, while later progesterone (increases in the second part of the menstrual cycle) to trigger the lobules in preparation for an impending pregnancy and these changes can be observed using magnetic resonance imaging (MRI)[185]. Breast cancer arises as a result of cellular deregulation of natural growth-reduction phases which typically leads to uncontrolled cell replication and tumour growth.

Breast cancer is classified into several (classical) subtypes based on tumour morphology according to behaviour (being either invasive or non-invasive) and location (ductal, lobular or metastatic) of the cancer[186,187]. For example, ductal carcinoma in situ (DCIS) and lobular carcinoma in situ (LCIS) are non-invasive cancers which are confined to the ducts and lobules of the breast, respectively[186,187]. In particular, location of the tumour is an important feature in patient prognosis[188].

It is not surprising that given the involvement of oestrogen and progesterone in the regulation of natural breast changes, that these hormones are also implicated in breast cancer development and progression. Furthermore, the use of HRT and hormonal birth control (specifically formulations used in 1960-1980) has consequently been associated with an increase in breast cancer incidence worldwide[29,189-192]. Since the early 1980s, it has been known that the effects of breast cancer related hormones are mediated by receptors like the oestrogen receptor (ER), progesterone receptor (PR) and human epidermal (growth factor) receptor 2 (HER2) and their status in a breast tumour can be used to further classify the disease[187,193-196]. Under this classification

scheme, breast cancers can be characterized according to the presence or absence of the hormone receptors, being the hormone positive (ER +ve and/or PR +ve), HER +ve, or triple negative breast cancer (TNBC; ER/PR/HER2 –ve)[187]. It is widely accepted that hormone receptor status is important in determining treatment strategy for patients and may also serve as a prognostic measure for treatment success, disease recurrence and overall survival (in non-metastatic breast cancers)[197,198].

In secondary or advanced breast cancer, the tumour has become invasive and spread to other parts of the body forming distant metastasis. This typically commences when breast tumour cells contaminate any of 30-50 axillary lymph nodes (though supra-clavicular, infra-clavicular and internal mammary nodes may also be involved) starting from level I nodes (or low axillar nodes, located in the lower armpit), followed by sequential progression to level II and level III nodes (or mid and high axillar nodes, located mid to high armpit and under the clavicles respectively)[199-203]. Once the lymphatic system is invaded, the cancer can readily spread to other regions of the body[204] typically to essential organs including the lungs, liver and brain as well as the bones. Metastasis typically results in the disruption of organ function and death is unable to be prevented in at this advanced stage of the disease.

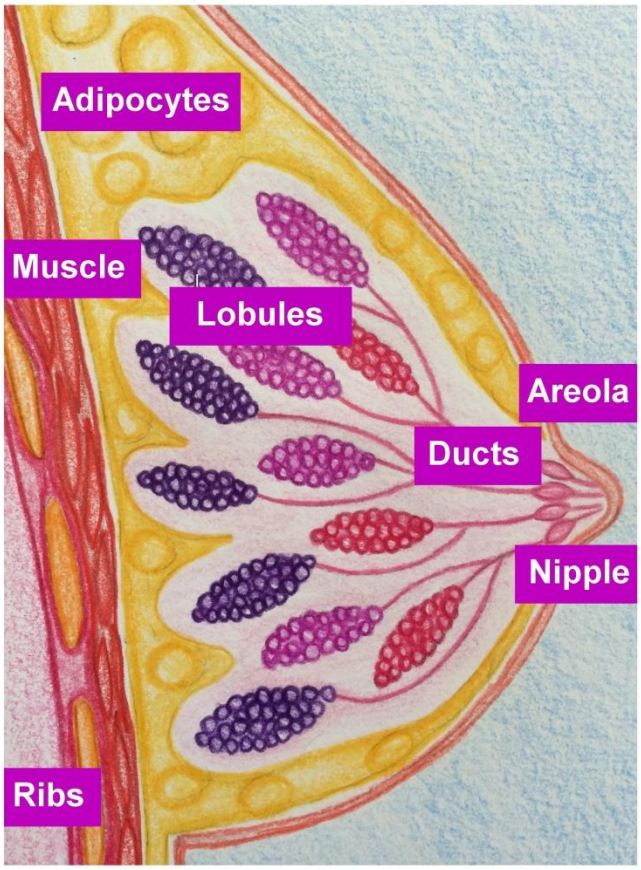


Figure 5 Illustration of breast structure.

The DSBR pathway

Like colorectal cancer, the genetic background of an individual in conjunction with exposure to environmental stimuli will influence an individual's likelihood of disease[205-208]. In fact, it's been reported that between 25-30% breast cancers could be prevented if healthy lifestyle choices were adopted by women and maintained throughout their entire life[209].

Breast cancer susceptibility can also be linked to an individual's ability to repair DNA damage, particularly DSBs. DSBs represent the most severe form of DNA damage and are associated with the breakage of both stands of the DNA sequence and this can result in chromosomal breaks or sequence rearrangements which on the cellular level can promote either apoptosis or tumour genesis[210]. Carcinogens (including chemotherapeutics and incidental chemical exposure), ionizing radiation, oxidative DNA damage, cell replication, programmed genomic rearrangements and meiotic DSBs are common agents and cellular processes which may result in DSBs[210].

The DSBR pathway has been the historical focus for identifying causative factors related to breast cancer development. This is because many of the genes known to be implicated in breast cancer risk, such as *BRCA1*, *BRCA2*, *ATM*, *TP53* and *CHEK2* all form part of these repair networks, being primarily HR and NHEJ next to less frequently described networks like micro-homology-mediated end joining (MMEJ) which is also known to occur[211-214]. Which network is utilized by a damaged cell will depend on which stage the cell is at in the cell cycle.

HR is considered to be a conservative and error free DNA repair pathway in which homologous DNA template is used as a template to regenerate missing sequence and effectively repairing the broken DNA strands (for a brief review on DSBR see[215]). When (1) cells become deficient in a gene protein (e.g. *BRCA1*) required for HR, (2) DNA damage arises outside S or G2 cell cycle phases (as cyclin-dependant-kinases are available that are required to promote end-resection) or (3) homologous DNA strands are not available as a repair template, cells are unable to undertake conservative DSBR via HR[216] and must thereafter rely upon NHEJ or MMEJ. Both these pathways are non-conservative error-prone repair methods which give rise to mutations in the form of duplications and deletions being incorporated into the DNA sequences[215]. Genetic alterations may thereafter accumulate in the genome leading to genetic instability and the development of malignancy.

Molecular testing of hereditary breast cancer

Familial breast cancer is estimated to account for 27% of all breast cancers and is associated with an earlier age of disease diagnosis and a higher frequency among family members[217,218]. Of these families 5-10% are suggested to harbor germline mutations (or complex genomic changes) that render inactive one of several high penetrance genes (including *BRCA1*, *BRCA2* and *TP53*) or moderate penetrance genes (*CHEK2*, *ATM* and *PALB2*)[217,219-221]. The largest contributors to breast cancer risk arise from mutations in *BRCA1* and *BRCA2* which have been reported to confer with a 65% and 45% relative risk of developing disease in mutation carriers[222] while lifetime risk of developing breast cancer in these women is estimated to be 60-85% and 40-85% respectively[217]. However mutations in these two genes are estimated to only account for only ~20% of patients[223,224] with approximately 7-10% of germline mutations being detected in *BRCA1* and about 10% in *BRCA2*[217] and as such a substantial number of breast cancer patients remain without a genetic diagnosis for their disease. Despite a plethora of genes and alleles conferring to increased breast cancer risk[221,225,226] having been reported in the past several decades and a non-subjective heightened presence of the disease in the affected families, the majority of breast cancer patients remain without a genetic diagnosis of disease. This suggests that either other genes (such as those residing in the DSBR pathway) or other mechanisms of gene inactivation of known breast cancer susceptibility genes may be responsible for disease in these patients.

Genetic Variation

No individuals have the same DNA. In identical twins the difference between DNA sequences can be detected even if they are as small as 0.1%[227-229]. These differences are referred to as genetic variation, and genetic variation is responsible for generating diversity within the species (variation in the genome translates to the variation observed in the phenotype)[230,231]. The evolution of a species relies on genetic variation as this provides the species with variation in traits and the ability to adapt to an ever-changing environment[232]. The primary source of genetic variation is gene flow, which can be observed through the migration of genes through a population or species over time as the variation is inherited from one generation to the next[231].

The vast majority of genetic variation is however unfavourable, and leads to either failure in reproduction, disease or the development of malignancy[1,233-235]. Genetic variation is primarily generated through incidental exposure to environmental stimuli in the form of DNA damage (somatic or non-inherited variants) due to the DNA being overwhelmed with mutagenic agents such as ROS, carcinogens, thermal disruption and radiation (i.e. UV light and x-rays)[236-240]. These agents (among others) are responsible for different types of DNA damage including: oxidation, alkylation, hydrolysis, bulky adduct formation, DSBs and mismatched bases[241,242]. Not all genetic variants are the same and certain types of genetic lesions are more likely to be generated in the presence of specific sources of DNA damage. For example oxidative stress, caused by exposure to free radicals can result in the development of 8-oxo-G lesions[243]; and ionizing radiation such as those created by x-rays, can generate damaged bases and abasic sites as well as single-strand and DSBs[244]. Consequently interactions between genetic and environmental factors contribute to disease development[20,35,206-209,245-248].

Furthermore, the development of specific types of cancer occurs in response to mutations arising in genes that are specific to the regulation of a given tissue (cancer susceptibility genes)[52,249]. The *MSH2* gene, for example, is expressed in epithelial cells and is responsible for the recognition of DNA mismatch errors and when functionally compromised can lead to cellular proliferation and ultimately malignancy, usually colorectal cancer when this damage occurs in the intestines[150-152].

Genetic variation may also become permanently incorporated into an individual's germline (inheritable DNA sequence)[1]. This type of genetic variation is established through the introduction of a *de novo* mutation (not present in the germline DNA of the

previous generation) that will be inherited in subsequent generations[250]. These variants are commonly associated with inherited disease or predispositions to disease development and arise at varying frequencies depending on the type of mutation incorporated[251,252]. Consequently, *de novo* instance of colorectal cancer have been suggested to account for up to 15% of all FAP patients and about 5% of all HNPCC patients[250,253,254].

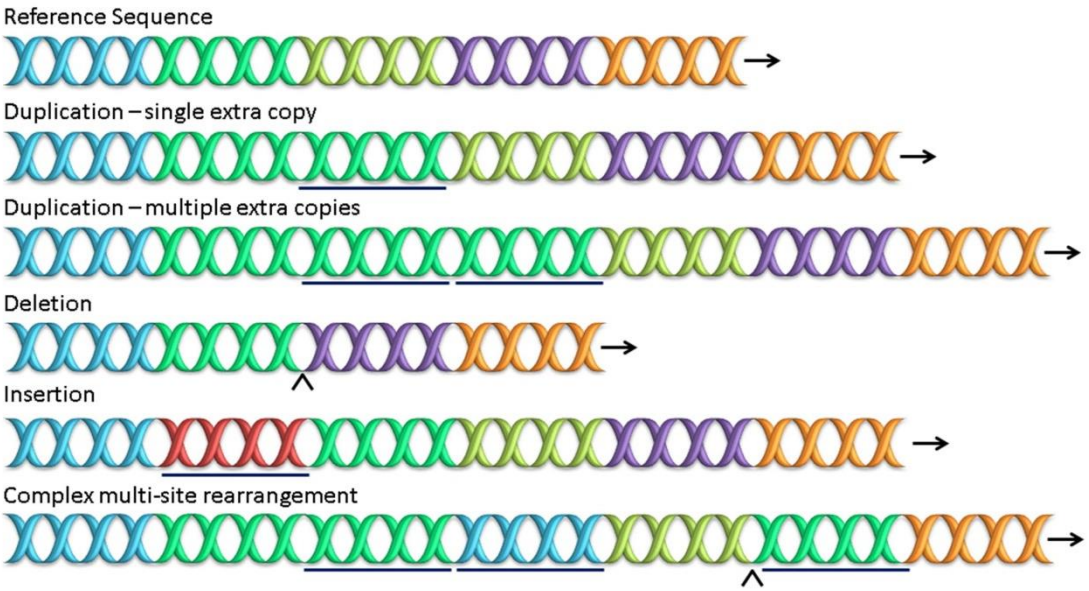
In relation to cancer development Knudson *et al.*[255] proposed the 'two-hit hypothesis', observing that individuals who carry mutations in cancer susceptibility genes harbor a change in one germline allele (predisposing to disease), but require a second "hit" to the second allele that is acquired later in life (somatic mutation) thereby resulting in loss of gene function and thus setting the scene for malignancy to develop. Cancer susceptibility genes and associated inherited germline variants have been identified for many common forms of cancer[256-261]. The Cancer Genome Atlas (TCGA), The Catalogue Of Somatically Acquired Mutations In Cancer (COSMIC) and the International Society for Gastrointestinal Hereditary Tumours (InSiGHT) are some of the databases available which provide in-depth information on cancers of both germline and somatically acquired origins and the genes and variants associated with them[262,263].

Genetic variation refers to a variety of molecular changes, and these are classified as either sequence variation or structural variation (see review[1]). DNA sequence variation describes genetic variation in the sequence of the genome. These sequence variants are classified according to the frequency at which they occur in the population. As such, polymorphisms, which are considered to have little or no influence on disease development, describe sequence variants occurring in over 1% of the population[264]. Deleterious sequence variants, or mutations, are found in less than 1% of the population and contribute greatly to the development of disease[265-267]. They describe variants such as inborn errors, germline mutations, or somatic mutations[1].

Structural variation

Unlike sequence variation, structural variation describes the microscopic and sub-microscopic sized variants, which alter the composition of the DNA. Structural variation describes changes that affect genomic regions where genes are located, the number of chromosomes, and how the DNA is folded into chromosomes[7,227]. Examples of structural variants include DNA tandem repeats, fragile sites, translocations, inversions, and CNVs[1]. As one of the most recently described forms of structural genetic variation, CNVs are yet to be characterized in most diseases including hereditary cancers[7,9-11] and therefore are the focus of this body of work.

Copy number (CN) is a term which describes the amount of copies of a particular region in the genome for which there are normally two copies. More or less copies, known as a CNV, may alternatively exist, indicating a gain or loss respectively, of genomic material. CNVs can be located in the DNA sequence of coding and regulatory regions of genes and hence they are considered to be of biomedical relevance[227,268,269]. Originally, CNVs referred to regions of DNA of one kilo base (Kb) or larger, however this has been redefined to include all DNA sequence variants with no minimum size[227,270]. CNVs describe genetic code variants such as duplications, insertions, deletions, and complex multi-site rearrangements[7,227,269] (see figure 6).



Direction of transcription indicated by arrow. Underlined regions or chevrons indicate the presence of a change that results in a new sequence that differs from reference sequence (top).

Figure 6 Illustration showing different types of CNVs

An understanding of how CNVs relate to the genetics of disease has increased as a result of the increased ability to undertake whole genome analysis. With the completion of the Human Genome Project in 2003, all genes in the human genome were identified and sequenced[271]. This project has led to the identification of genes associated with disease and discoveries into the modes of disease transmission, assessment of disease risk and disease modification[227,268,270,272]. The discovery of CNVs has since lead to the creation of the first CNV map of the human genome (completed using the HapMap collection) which revealed dramatic variation in the CN of genomic regions within populations[9-11,273]. The study also observed numerous examples of CNVs located in genomic areas with possible medical relevance, estimating that 15.6% of all genes overlap with reported CNV sites[273]. Several studies of CNVs in human cohorts has also identified a widespread presence of CNVs in healthy individuals[10,268-270,274-277]. CNVs are noted to span many thousands of bp, often encompassing and disrupting functional DNA sequences resulting in disease[278-281]. A study by Girirajan and Eichler[282] has furthermore suggested that the severity of disease may be explained as a result of the overall burden of CNVs in an individual's genome. They proposed that disease risk is correlated with increased CNV burden and that variation in CNV burden will result in variation in disease phenotype[282].

It is considered that complex diseases are associated with changes in gene expression and that any form of genetic variation, including CNV changes, may be involved in the disease process[12]. CNVs have been implicated in the development of various forms of cancer, including inherited colorectal cancer syndromes like LS, FAP and JP as well as breast cancer[84,106,112,283-294]. Numerous studies have also identified various mechanisms in which CNVs can result in disease, including: directly disrupting functional gene sequence, alu-mediated recombination, alternative splicing, paracentric inversions, and promoter region inactivation[84,106,112,283,284,286-289,295-297].

CNVs have been reported to alter gene expression, phenotypic variation in disease and adaptation through directly influencing gene dose[298], indicating a likely involvement in disease development. Altered gene dose occurs when the abundance of a particular sequence within the genome (i.e. through loss or gain of genetic material) results in changes to the quantity of the expressed gene transcript (i.e. increased or decreased gene expression). CNVs can influence gene dose and they can result in changes to epigenetic patterning, cause disruption to microRNA (miR) controlling species, and via the disruption of non-coding gene sequences.

Epigenetic patterning and CNVs

Epigenetics describes a variety of genetic changes which alter gene expression without changing the DNA sequence, and these changes can be inherited in a process called epigenetic memory[299]. Epigenetic changes include DNA methylation, histone modification and small RNA changes and these are of particular interest in cancer as they can be targeted by chemotherapeutic agents[300-302]. Environmental factors such as diet are known to remodel epigenetic patterning and since this discovery, significant research has commenced into epigenetic therapies for cancer prevention and treatment[302-306].

CNVs have been implicated in disease development by means of generating aberrant epigenetic regulation in cells[70-72,144,307]. Epigenetic modification is known to contribute to sporadic colorectal cancer and recent evidence also suggests a role in hereditary colorectal cancer predispositions[307,308]. Ligtenberg *et al.*[144] identified a deletion in the 3' region of *EPCAM* containing the polyadenylation signal whereby loss of this region resulted in transcription read-through of *MSH2*. Studies have further delineated that other variants in the *EPCAM* gene result in epi-mutations associated with a preferential risk of endometrial cancer[70-72,307]. Furthermore, a CN gain has also been reported to be associated with the methylation of *MLH1* in some LS families[295,307]. With regards to breast cancer, *Birgisdottir et al.*[309] have reported the methylation of *BRCA1* in the presence of a *BRCA1* deletion, to be a frequent event in sporadic breast tumours.

MiR expression and CNVs

MiRs are small non-coding RNAs (~22 nucleotides in length) which pair to the 3' un-translated regions of messenger RNAs (mRNAs) preventing them from being translated in to protein (for a full review see[310])[311-313]. MiRs contribute to tumorigenesis by acting as tumour suppressors or tumour promoters[314], where loss of miR expression can result in the over-expression of its target gene(s), and gain of miR expression can result in the under expression of its target gene(s). Given that the expression of hundreds of genes may be influenced by a single miR (including those that influence cell adherence, migration, invasion, motility and angiogenesis etc), research into miRs is a rapidly growing area of cancer interest[315-318].

The majority of miR studies thus far have focused on characterising altered levels of miR expression in disease[319-322]. Motoyama *et al.*[323] reported *mir-31*, *mir-183*, *mir-17-5*, *mir-18a*, *mir-20a* and *mir-92* to be down-regulated, and *mir-143* and *mir-145* up-regulated in colorectal cancer and have suggested down-regulation of *mir-18* to be

associated with poor patient prognosis[323]; Landi *et al.*[324] reported polymorphisms in miR binding sites that may be associated with an increased risk of colorectal cancer. Similarly, miRs have also been identified to contribute to breast cancer development (reviewed in[325]). While many of these studies need further validation and functional studies performed to determine the role of miR changes in disease development and progression, collectively they suggest aberrant miR disease profiles are associated with the underlying disease processes. CNVs have emerged as a mechanism of miR inactivation causing disease[326]. A study by Zhang *et al.*[327] has characterized CNVs in melanoma, ovarian and breast cancers, reporting miR CN changes correlate with changes in miR expression which are suggested to impact on target gene expression resulting in the development of disease.

Non-coding gene regions and CNVs

Non-coding regions of genes describe any genetic sequence which is not transcribed into mRNA, and includes DNA sequences belonging to the intronic or promoter regions of a gene. During transcription, in a process called RNA splicing, introns are removed from between coding exons to generate mature RNA products[328,329]. Aberrations residing within intronic regions have the capability of altering the site in which break-points between introns and exons occur, leading to a variety of aberrations including translocation-gene-fusions, as seen in some cases of acute lymphoblastic leukaemia due to a breakpoint between exons 1 and 2 fusing the *BCR* and *ABL* genes[330,331]. Intronic variants can also generate cryptic splice sites which may result in the inclusion of additional sequence (pseudo-exon) into the mature RNA product potentially affecting its function (for a full review of the mechanisms see[332])[333,334].

Variation within the intronic regions of genes is recognized to be a novel mechanism for disease development in hereditary breast cancer and hereditary colorectal cancers[296,335-339]. Most recently, an intronic deletion 478 bp upstream of exon 2 of *MSH2* was identified as the cause of LS[296], while a 1.4 Kb deletion within intron 14 of the *APC* gene is known to be the cause of FAP[335]. A 250 Kb deletion from intron 5 to exon 15 and a substitution at the splice donor site of intron 9, have also been described in the *APC* gene and were associated with the development of FAP in two respective families[336].

It is important to recognise that CNVs located in and around promoter regions of genes can contribute to disease as structural alterations in this region can can prevent the binding, recruiting or activation of transcription factors which is required to initiate RNA

polymerase activity (for transcription to occur)[106,107,112,295,340]. The promoter region of a gene is an important regulatory region and is comprised of a specific DNA sequence which allows transcription factors and RNA polymerase to bind and commence the synthesis of the RNA. The promoter region is typically located upstream at the 5' end of the gene in which many Kb of DNA sequence may be present before the transcription start site. Additionally, multiple promoter regions for one gene may exist, which makes the study of promoter regions more complex. For example, recent studies have identified deletions encompassing the promoter 1A and 1B regions of the *APC* gene to result in the silencing of the gene and the cause of FAP[106-108,112].

Aims and Hypothesis

Aims

1. To identify CNVs on a genome wide level using a Cytogenetic Whole Genome array in a series of hereditary breast cancer, hereditary colorectal cancer and healthy control genomes.
2. Conduct a CNV association analysis using genome-wide data to identify novel or highly significant CNVs between patients and control genomes.
3. Perform targeted CNV screening for duplications and deletions residing in and in vicinity of known cancer susceptibility genes and genes associated with cancer susceptibility pathways.
4. Carry out pathway analysis for each of the patient gene catalogues revealed from the CNV analysis in the aim of uncovering biologically meaningful relationships that may underpin disease development.
5. Undertake mRNA transcript analysis for *MLH1*, *MSH2* and *MSH6* in a cohort of HNPCC patients to investigate the contribution of deep intronic variants in disease development.

Hypothesis

Copy number variants (CNVs) are associated with the development of disease in a proportion of families with a clinical diagnosis of hereditary breast cancer, hereditary non-polyposis colorectal cancer (HNPCC) or familial adenomatous polyposis (FAP) where no mutations in genes known to be associated with their disease (*BRCA1*, *BRCA2*, *MLH1*, *MSH2*, *MSH6*, *PMS2, APC* or *MUTYH*) was identified.

CHAPTER 2: COPY NUMBER VARIATION IN HEREDITARY POLYPOSIS

Introduction

Familial adenomatous polyposis (FAP) is the second most common inherited predisposition to colorectal cancer and is associated with the development of hundreds to thousands of adenomas in the colon and rectum[90]. Average age of colorectal cancer diagnosis is ~36 years[21,76,89,90].

Somewhere between 80% and 90% of all patients seeking genetic testing for FAP, mutations in either *APC* or *MUTYH* will be identified[94,100].. Since their discovery, over 1500 mutations have been identified in *APC* and a further 300 have been identified in *MUTYH* [92,114,115,341]. For the remaining 10-20% of patients no genetic diagnosis can be identified suggesting other genes or mechanisms that render *APC* or *MUTYH* inactive may be responsible for disease in these polyposis patients.

Copy number variants (CNVs) represent a form of structural genetic variation associated with a gain or loss of genomic material which have been shown to contribute to disease development[144,286,288,295,296,342]. CNVs remain to be investigated in hereditary polyposis and may account for proportion of patients where no genetic diagnosis has been found.

This part of the thesis aims to describe CNVs identified in the genomes of patients diagnosed with FAP who do not harbour germline mutations in *APC* or *MUTYH*. Furthermore, CNVs identified will be compared to a cohort of control genomes and the Database of Genomic Variants (DGV) thus enabling the identification of unique and rare CNVs, respectfully, which may be involved in the pathogenesis of hereditary polyposis.

Publication

Amy L. Masson, Bente A. Talseth-Palmer, Tiffany-Jane Evans, Patrick McElduff, Allan D. Spigelman, Garry N. Hannan and Rodney J. Scott (2015) Copy number variants associated with 18p11.32, *DCC* and the promoter 1B region of *APC* in familial adenomatous polyposis, *Gene*.

Co-author statement

I attest that Research Higher Degree candidate Amy Louise Masson contributed to the above manuscript including involvement in the conception and design of the study; conducting the laboratory work, data analysis and interpretation; and preparation of the manuscript.

Co-author	Signature	Date
Bente A. Talseth-Palmer		
Tiffany-Jane Evans		
Patrick McElduff		
Allan D. Spigelman		
Garry N. Hannan		
Rodney J. Scott		

Amy Louise Masson Date: 01/09/2015

Professor Robert Callister Date: 01/09/2015

Assistant Dean Research Training

Copy number variants associated with 18p11.32, *DCC* and the promoter 1B region of *APC* in colorectal polyposis patients

Amy L. Masson[1,2], Bente A. Talseth-Palmer[1,2], Tiffany-Jane Evans[1,2], Patrick McElduff[3,4], Allan D. Spigelman[5,6,7], Garry N. Hannan[8] and Rodney J. Scott[1,2,9]

[1]Centre for Information-Based Medicine, Hunter Medical Research Institute, University of Newcastle, Newcastle, New South Wales, 2305 Australia;

[2]School of Biomedical Sciences and Pharmacy, Faculty of Health and Medicine, University of Newcastle, New South Wales, 2308 Australia;

[3]Centre for Public Health, Hunter Medical Research Institute, University of Newcastle, Newcastle, New South Wales, 2305 Australia;

[4]School of Medicine and Public Health, Faculty of Health and Medicine, University of Newcastle, New South Wales, 2308 Australia;

[5]Hunter Family Cancer Service, Hunter New England Area Health, Newcastle, New South Wales, 2305 Australia

[6]University of NSW, St Vincent's Hospital Clinical School, Sydney, New South Wales, 2010 Australia

[7]Hereditary Cancer Clinic, St Vincent's Hospital, The Kinghorn Cancer Centre, Sydney, New South Wales, 2010 Australia

[8]CSIRO Food and Nutrition Flagship, North Ryde, New South Wales, 2113 Australia;

[9]Division of Molecular Medicine, Hunter Area Pathology Service, John Hunter Hospital, Newcastle, New South Wales, 2305 Australia;

Correspondence to

Professor Rodney J. Scott

Information Based Medicine Program, Hunter Medical Research Institute

Level 3 West, New Lambton Heights, New South Wales, 2305 Australia

Email: Rodney.Scott@newcastle.edu.au Tel: +61 (2) 4921 4974

Abstract

Familial Adenomatous Polyposis (FAP) is the second most common inherited predisposition to colorectal cancer (CRC) associated with the development of hundreds to thousands of adenomas in the colon and rectum. Mutations in *APC* are found in ~80% polyposis patients with FAP. In the remaining 20% no genetic diagnosis can be provided suggesting other genes or mechanisms that render *APC* inactive may be responsible. Copy number variants (CNVs) remain to be investigated in FAP and may account for disease in a proportion of polyposis patients. A cohort of 56 polyposis patients and 40 controls were screened for CNVs using the 2.7M microarray (Affymetrix) with data analysed using ChAS (Affymetrix). A total of 142 CNVs were identified unique to the polyposis cohort suggesting their involvement in CRC risk. We specifically identified CNVs in four unrelated polyposis patients among CRC susceptibility genes *APC*, *DCC*, *MLH1* and *CTNNB1* which are likely to have contributed to disease development in these patients. A recurrent deletion was observed at position 18p11.32 in 9% of the patients screened that was of particular interest. Further investigation is necessary to fully understand the role of these variants in CRC risk given the high prevalence among the patients screened.

Key Words

cancer, polyposis, CNV, long non-coding RNAs, diagnostic testing

Introduction

FAP is an autosomal dominant inherited disease, which affects nearly 1 in 12,000 individuals and accounts for approximately 0.5% of all CRCs. Typically, FAP is characterized by the early development of hundreds to thousands of adenomas in the colon and rectum. The development of adenomas commences during early childhood and adolescence and commonly becomes malignant if untreated, with an average age of cancer onset of 35-36 years[1]. A less severe form of FAP termed attenuated FAP is characterized by fewer adenomas and a later age of disease onset[2,3].

Mutations in the *APC* gene were found to be the genetic basis of FAP and since its discovery over 1500 pathogenic mutations have been reported[4,5]. Up to 20% of polyposis patients do not have a family history of disease but do harbour germline *APC* mutations. Mutations in the *APC* gene account for the majority of patients diagnosed with FAP and more recently, mutations in the base excision repair gene *MUTYH* have been shown to be associated with a recessive form of colorectal polyposis[6]. Up to 20% polyposis patients that are clinically tested for mutations in these genes do not have a germline mutation and no genetic diagnosis for their disease.

CNVs represent a form of structural genetic variation associated with a gain or loss of genomic material. CNVs have been shown to contribute to the development of disease directly through the disruption of functional gene sequences; via promoter region inactivation; or as a result of more cryptic changes such as alterations in epigenetic marks, changes to microRNA controlling species, transcription read through, unmasking recessive alleles and via disruption of non-coding gene sequences[7-12].

Furthermore, while CNVs which are commonly observed in the population may contain cancer related genes, it is the rare CNVs (low population frequencies) which are proposed to harbour genes or other regulatory elements that are likely to be disease susceptibility factors[13]. Several studies have recently investigated the contribution of rare CNVs in cancer; one study identified 26 rare CNVs which they proposed to contribute to breast cancer susceptibility, while another has reported the enrichment of disrupted genes that affect the maintenance of genomic integrity i.e. DNA double-strand break repair also in familial breast cancer[14,15].

In this study we have focused on the role of CNVs in the genomes of patients diagnosed with polyposis that do not harbour germline mutations in *APC* or *MUTYH* as assessed by direct DNA sequencing and multiplex ligation probe amplification (MLPA). High throughput microarray technology has continuously improved since its

introduction such that now continually smaller CNVs can be detected in ever larger patient cohorts. We used the Affymetrix Cyto2.7M microarray, which at the time of this study provided the highest genomic coverage of any commercially available microarray; containing 400,000 SNP probes and >2.1 million CNV probes with an average spacing of 1395 base pairs (bp). CNV analysis was conducted on DNA derived from 56 polyposis patients (*APC*/*MUTYH* mutation negative) and compared to 40 controls and the Database of Genomic Variants (DGV) with the aim of identifying CNVs, which may be involved in the pathogenesis of the observed disease.

Methods

Samples

The study including patient recruitment and all experimental protocols were approved by the Hunter New England Human Research Ethics Committee and the University of Newcastle Human Research Ethics Committee. The methods employed in this study were carried out in accordance with the approved guidelines of the University of Newcastle. Genomic DNAs were obtained from polyposis patients who had given informed consent for their DNA to be used for studies into their disease and control DNA samples from the Hunter Community Study was used in the current study[16]. DNA was extracted from whole blood by the salt precipitation method[17].

The inclusion criteria for this study was a patient diagnosed with adenomatous polyposis or and who did not have a detectable *APC* or *MUTYH* mutation as assessed by complete Sanger sequencing and MLPA analysis. A cohort of 56 clinically histologically confirmed polyposis patients was used in this study. All patients were unrelated and were diagnosed after colorectal resection who then sought genetic testing for their condition. The average age of diagnosis was 51 years (range 10 - 74), 32 of the probands had a family history of colonic polyposis or CRC, 21 had no family history and for 3 patients no information on family history was available. Polyp counts ranged from 5 through to over 1000, however most patients had less than 100 polyps suggesting that the majority of patients presented with an attenuated form of polyposis. Genomic DNA from 40 unrelated healthy (and not affected with any cancer during their life) subjects who were >55 years at the time of sample collection were available as controls. A total of 96 samples were included in the study.

Genomic array preparation and data processing

The genomic DNA was processed on the Affymetrix Cyto2.7M array according to manufacturer's protocols. Affymetrix Chromosome Analysis Suite (ChAS) (Version CytoB-N1.2.0.232; r4280) was used to analyse the array data (NetAffx Build 30.2 (Hg18) annotation).

A training set of 20 randomly selected samples was used to further optimize a series of quality control (QC) parameters reduce the number of false-positive CNVs being included in the analysis as many of these QC thresholds were more stringent than default settings alone.

The Cyto2.7M array is comprised of both CNV probes and SNP probes. For the confident detection of CNVs samples were required to have a minimum quality threshold of: mapdQC <0.27 (Median Absolute Pair-wise Difference; QC of CN probes compared to a reference model); snpQC >1.1 (SNP probe QC measuring distances between the distribution of alleles AA, AB and BB alleles in which larger differences in allele distribution is associated with an increased ability to call a given genotype); and wavinessSd <0.1 (measure of standard deviation in data waviness; the GC content across the genome correlates with average probe intensities).

For the data that fulfilled the array performance QC, CNV calling QC was then undertaken to minimise the inclusion of false positive or negative CNV calls being incorporated into the analysis. This included evaluation of CNV calls with respect to having >90% confidence, a CNV having to be of autosomal origin (not located on either sex chromosome), and CNVs had to have a minimum of 24 probes used to call the CNV region. Visual inspection was used to confirm all CNV calls, verify the suggested CN state, and to further identify regions associated with low marker coverage and excluded across all samples (i.e. centromeric and telomeric regions; see supplementary table 7). The smallest CNV detected with confidence was 6.03 Kilobases (Kb) across all samples.

CNV and statistical analysis

CNVs were subject to a series of analyses which included: (1) Identification of abundant genomic regions and genes affected by a CNV in patients; (2) Statistical assessment of the distribution of CNVs across the genome of patients compared to controls; and (3) interrogation of CNV data for CN gains and losses residing in or ±100 Kb of 77 genes comprising the WNT signalling and mismatch repair (MMR) pathways as well as other reported CRC susceptibility genes, likely to be associated with polyposis (see supplementary table 4)[6,18]. Associations (e.g. numbers and sizes of CNVs) were statistically compared between patients and controls using a two tailed un-paired t-test Graphpad Prism (Version 6; available http://www.graphpad.com/quickcalcs/ttest1/). The derived *p-values* were corrected for multiple testing using Bonferroni correction (Alpha=0.05, R= 22). The Bonferroni's adjustment resulted in the confidence level being <0.0023.

Validation of CNV results

CN gains and losses were subject to validation using pre-designed TaqMan Copy Number Assays (Applied Biosystems). Where possible, two CN assays (test assays)

located within and two located just outside (control assay) the CNV of interest were utilized (assay information summarized in supplementary table 8). The sample(s) of interest were tested along with a no template control (NTC) and several calibrator samples (of known CN for the region assessed). All samples were assayed in triplicate and real-time PCR was conducted according to manufacturer's instructions using 10 ng of DNA sample in a final reaction volume of 20 μL. Using the real-time PCR (Applied Biosystems 7500; SDS software Version v1.4) the assay was run according manufacturer's protocols. CopyCaller v2.0 software (Applied Biosystems) was used to analyse the results.

Several CNVs were further validated using this independent method (see supplementary table 9). These CNVs included both CN gains and CN losses in the genes *DCC* and *APC* as well as in the genomic region 18p11.32. As we observed high concordance between array data and all CNVs were validated using an independent method, it was considered unnecessary to confirm every CNV identified as analysis parameters were consistent across all samples.

Pathway analysis and annotation

In silico analysis conducted in this study involved the analysis of 49 of the 148 genes unique to the patients that were also considered rare as they have not been reported in the DGV).

Pathway analysis was performed using WebGestalt software (Version 2013)[19]. This software was used to assess gene lists derived from the refined CNV results obtained from ChAS according to Kyoto Encyclopaedia of Genes and Genomes (KEGG) pathways, cytoband enrichment analysis, and miR targets. Analysis was performed using hypergeometric statistical method, Benjamini and Hochberg (BH) correction for multiple testing and a biological significance threshold of <0.05 with a minimum of two genes per category required to assess any enrichment.

TAM (Tool for Annotations of miRs) (Version 2)[20] software was used to annotate miRs according to miR family, cluster, function, Human miR associated disease database (HMDD) and tissue specificity. Annotations were performed using the following parameters: all miRs in the TAM database were used as a background; to identify meaningful categories we looked at miR over-representation in all categories and analysis was limited to at least one miR in a given category. Enrichment analysis for miRs categories was conducted using hypergeometric testing and *p*-values were corrected according to Bonferroni correction for multiple testing.

Results

Array resolution and CNV detection

A total of 278 CNVs were identified in the 96 participants involved in this study (table 1). CNVs ranged in size from 6.03 Kb to 1435.95 Kb. The average number of CNVs identified per sample did not differ significantly between patients and controls (*p=0.4383*) nor did the average CNV burden (*p=0.5173*) or average CNV size (*p=0.1664*).

Table 1 Summary of CNV results obtained from the Cyto2.7M array analysed in ChAS.

		CNV Count		CNV Size (Kb)		
		Median CNVs per sample	Mean CNVs per sample	Total CNV affected genome per group	Mean total CNV affected genome per sample	Mean size of a CNV
Patients	56	2	3.11	14,018.83	250.34	82.18
Controls	40	2	2.6	11,820.75	295.52	106.57
p	-	-	*0.4383*	-	*0.5173*	*0.1664*

*statistically significant

Abundant genomic regions and genes associated with CNVs

Analysis of the control population revealed a total of 104 CNVs of which 12 genomic regions were disrupted by a CNV in more than one individual. Eight of the genomic regions (2p16.1, 4p15.31, 4q13.1, 5p13.3, 5q21.2, 7p14.1, 8q12.1 and 8q24.23) were disrupted by a CNV in two unrelated individuals; three genomic regions (4q32.2, 6q22.31 and 12p13.31) were disrupted by a CNV in three control participants; and one genomic region (3q26.31) was found to be affected in five individuals (see supplementary table 1). In total 66 of the 104 CNVs (63.46%) disrupted 96 genes. The 96 genes disrupted by CNVs were screened against the current cancer genome census list (COSMIC database available: http://cancer.sanger.ac.uk/cancergenome/projects/cosmic/) to determine whether any were likely to be associated with cancer. None of these genes were in common with those known to be associated with a cancer predisposition. Three CNVs were identified to disrupt *MLL3*, *PBX1* and *PLAG1,* respectively, that have been observed in medulloblastomas, pre-BALL/myoepitheliomas and salivary adenomas.

Of 174 CNVs identified in the polyposis patients, 32 contained genomic regions that were common to those identified in controls (see supplementary table 2). These CNVs were not included in further analysis as they were considered most likely to be neutral. Of the remaining 142 CNVs unique to the patients, 6 genomic regions contained CNVs which were common to multiple patients (table 2). The genomic region 2q32.3, was disrupted by a CNV in two patients, as were the CNVs that encompassed regions located at 2q34, 3q26.1 and 4q12; one genomic region (3q26.32) was disrupted by a CNV in three patients; and another genomic region (18p11.32) was disrupted by a CNV in five patients.

Table 2 Recurrent CNVs unique to polyposis patients. Note location of recurrent CNV, the type of CNV identified in each patient, description of the CNV (Chr, Start, End, size) and patient ID.

Location	CNV Type	Chr	Start (bp)	End (bp)	Size (Kb)	Patient IDs
2q34	Gain	2	209,751,678	209,923,081	171.4	FAP6
	Gain*	2	209,793,755	209,821,161	27.41	FAP23
2q32.3	Loss	2	194,626,162	194,695,204	69.04	FAP19
	Loss	2	194,626,162	194,696,112	69.95	FAP20
3q26.1	Loss	3	166,523,809	166,565,186	41.38	FAP4
	Loss	3	166,523,809	166,565,186	41.38	FAP17
4q12	Gain	4	57,745,642	57,794,798	49.16	FAP24
	Gain	4	57,745,642	57,794,798	49.16	FAP4
3q26.32	Loss	3	177,370,126	177,396,832	26.71	FAP21
	Loss	3	177,370,126	177,396,832	26.71	FAP5
	Loss	3	177,370,126	177,396,832	26.71	FAP22
18p11.32	Loss	18	1,891,809	1,974,284	82.48	FAP4
	Loss	18	1,894,368	1,974,284	79.92	FAP1
	Loss	18	1,894,368	1,974,284	79.92	FAP2
	Loss	18	1,894,368	1,974,284	79.92	FAP3
	Loss	18	1,964,144	2,015,983	51.84	FAP5

*indicates the CNVs not reported in the DGV

In three patients CNVs located 6p12.3, 20p12.1 and 22q12, respectively, harboured large deletions encompassing part of the genes *SUPT3H*, *MACROD2* and *SYN3* and were subject to special consideration as the three genes were also affected by CNV changes in 3 control subjects (see figure 1). Two of the deletions occurring in the patients encompassed coding regions of *SUPT3H* (exons 4-10) and *SYN3* (exons 7-9) suggesting a loss of gene function in both instances. The *SUPT3H* CNV in another control subject occurred in intron 2 and appears to be less likely to affect function. The *SYN3* CNV in a control subject was a duplication that included exons 4 and 5 and which did not alter the reading frame of *SYN3* and may have affected gene function. The CNV affecting *MACROD2* in intron 5 of a polyposis patient appeared not to alter exonic structure of the gene. Similarly, a control subject was also found to harbour a CNV in intron 5 of *MACROD2* but residing 5' to that identified in the patient. Neither CNV encompassed an exon. Since two adjacent deletions in intron 5 of *MACROD2* were observed that did not appear to alter the exonic structure of the gene they were not considered to be disruptive.

Of the remaining 139 CNVs (see supplementary table 3), 85 (61.15%) disrupted a total of 148 genes and these were considered candidate genes for disease development in these patients. Furthermore a subgroup of 10 genes: *EVI2B*, *EVI2A*, *SMAP2*, *BOD1L*, *NAMPT*, *NF1*, *HSD11B1*, *G0S2*, *DOCK4* and *A2BP1* were found to be affected by a CNV in more than one patient (table 3) and therefore were considered to have a higher probability of being associated with disease warranting further investigation.

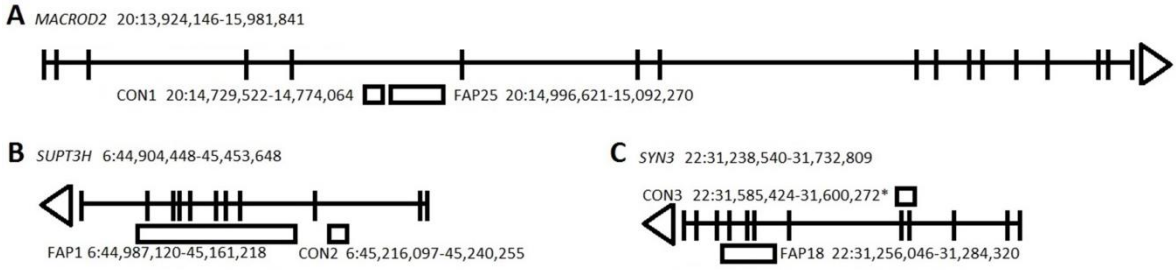


Figure 1 Genes within which non-overlaping CNVs were identified in patients and controls: (A) *MACROD2*, (B) *SUPT3H* and (C) *SYN3* respectively.

Note the location of the CNVs (duplication above and deletions below) with respect to the gene identifying exons and introns and direction of transcription (direction of arrow). Representation only, not to scale.

Table 3 The 10 genes associated with CNVs unique to polyposis patients (identified as FAP11-16). Note the disrupted gene (its symbol and description), if the gene is expressed in the colon (www.proteinatlas.org), the CN type observed in the current dataset and a general column outlining the predicted interpretation of the effect different types of CNVs have on disease development.

Gene	Description	Expression in colon	CN type	Region of gene disputed	Interpretation (predicted)
EVI2B	ecotropic viral integration site 2B	medium	Gains	*FAP11: whole gene* *FAP12: part gene*	*Whole gene duplication: increased expression/amplification of gene function.*
EVI2A	ecotropic viral integration site 2A	medium	Gains	*FAP11: whole gene* *FAP12: whole gene*	
SMAP2	small ArfGAP2	medium	Gains	*FAP12: part gene, exon and intron 1* *FAP13: part gene , exon and intron 1*	*Partial duplication or deletion involving introns and exons: disruption of gene/loss of gene function.*
BOD1L	biorientation of chromosomes in cell division 1-like 1	medium	Gains	*FAP12: part gene, introns and exons* *FAP11: part gene, introns and exons*	
NAMPT	nicotinamide phosphoribosyltransfer-ase	medium	Gains	*FAP12: part gene, most of it from start of gene* *FAP13: whole gene*	*Partial duplication involving promoter: transcription of aberrant transcript leading to non-functional protein product/loss of gene function.*
NF1	Neuro-fibromin 1	medium	Gains	*FAP11: part gene, intronic* *FAP12: part gene, intronic*	
HSD11B1	Hydroxyl-steroid (11-beta) dehydrog-enase 1	low	Gains	*FAP12: part gene, upstream into exon 1* *FAP11: part gene, upstream*	*Partial duplication involving exons: possible*

				into exon 1	*addition of duplicated exons into gene transcript creating a non-functional protein product/loss of gene function.*
G0S2	G0/G1 switch 2	medium	Gains	*FAP12: whole gene* *FAP11: whole gene*	
DOCK4	dedicator of cytokinesis 4	low	Both	*FAP11: part gene dup, intronic* *FAP14: part gene del, intronic*	*<u>Partial duplication or deletion involving introns:</u> development of cryptic splice sites or the formation of pseudo exons leading to disruption in gene expression/loss of gene function.*
A2BP1	RNA binding protein, fox-1 homolog (C. elegans) 1 (alias *RBFOX1*)	low	Both	*FAP15: part gene dup, introns and exon* *FAP16: part gene del, introns and exons*	

Distribution of CNVs across the genome in patients

The distribution of CNVs across chromosomes was compared between patients and controls revealing no statistically significant differences in CNV distribution. We did observe a trend in the over-representation of CNVs in patients (CNVs=9) compared to controls (CNVs=0) using Fisher's exact test for chromosome 18 (p=0.009) which did not remain statistically significant after correction for multiple testing (i.e. p>0.0022).

Among the CNVs located on chromosome 18, a CNV gain was detected in the first intron of the *DCC* gene (18:48,381,778-48,412,417; 30.6 Kb; 93% confidence, detected by 52 probes), which was subsequently confirmed by TaqMan CN assays (see figure 2).

A CNV gain was also identified in one patient which encompassed exon 1 and extending into the first intron of the *USP14* gene located on chromosome 18 (position 18:50,739-154,914, size 104 Kb; 90% confidence and detected by 34 probes). This CNV encompasses part of the irritable bowel disease (IBD) locus[21] and may have contributed to the disease phenotype of this patient.

Of note, five patients harboured the same CNV loss at the 18p11.32 locus (figure 3). The predicted size of the largest CNV identified a loss of 82.48 Kb (93% confidence and detected by 65 probes); three other unrelated patients all harboured a similar sized CNV loss of 79.92 Kb (all >91% confidence and detected by 63 probes); and one patient was found to have a CNV loss of 51.84 Kb (94% confidence and detected by 59 probes). All CNVs located in the 18p11.32 region overlapped each other by 10.14 Kb and this was confirmed by TaqMan CN assay to be lost in all five patients. This region of loss is proposed to contain the long non-coding RNA (lnc-RNA) TCONS_00026231 (18:1,963,908-1,972,876 Hg19; UCSC Genome Browser).

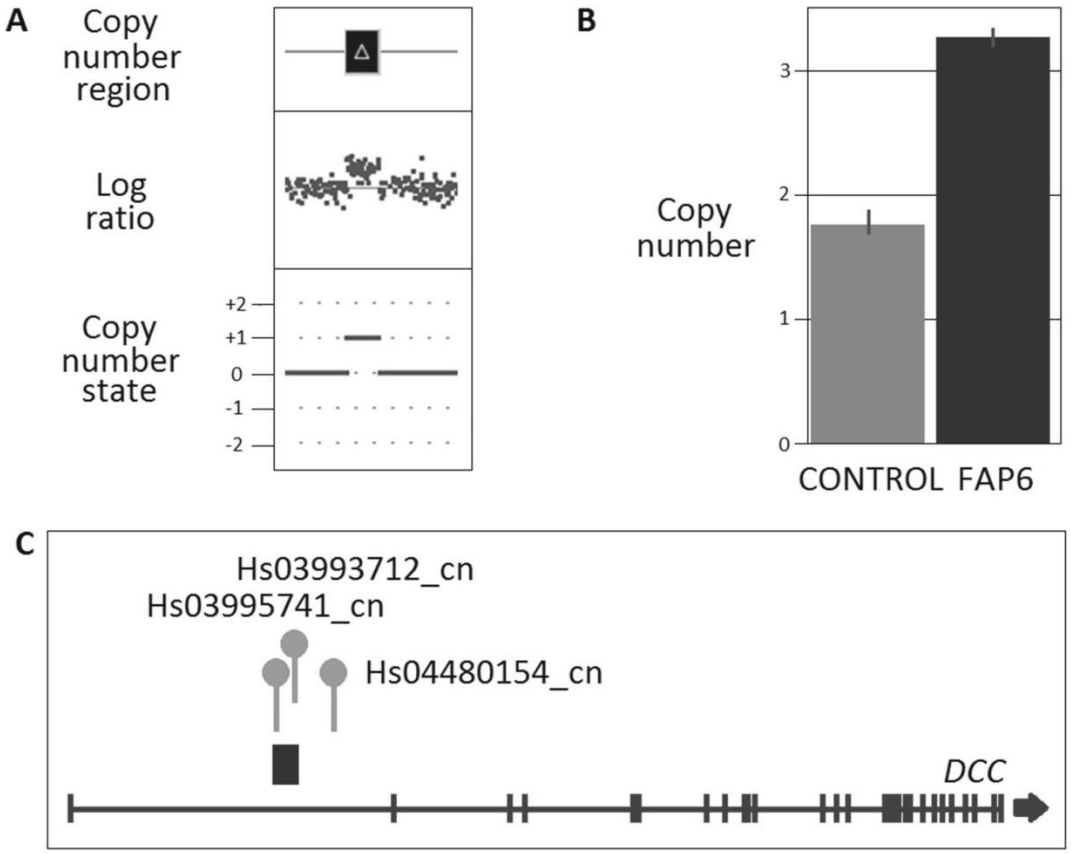


Figure 2 CNV results for the duplication in the *DCC* gene in an FAP patient (FAP6). (A) CNV profile from Cyto2.7M array data using ChAS noting the defined CN region (dark box above gene representing the CN deletion) in relation to the log ratio plot (relative fluorescence of each probe, dot, on the array showing a decrease in fluorescence indicating a loss in genomic material) and the CN state (0= normal two copies present, +1= one extra copy, +2= two extra copies, -1=one less copy and -2= two less copies); (B) Validation using TaqMan CN assay showing results for assay Hs03995741_cn noting the normal two copies of this region identified in the control (CON1), confirmation of the aberrant three copy in the affected FAP patient and the error bars associated with the three technical repeats for each sample; and (C) Location of CN duplication with respect to the gene and the TaqMan CN assays used in validating the variants.

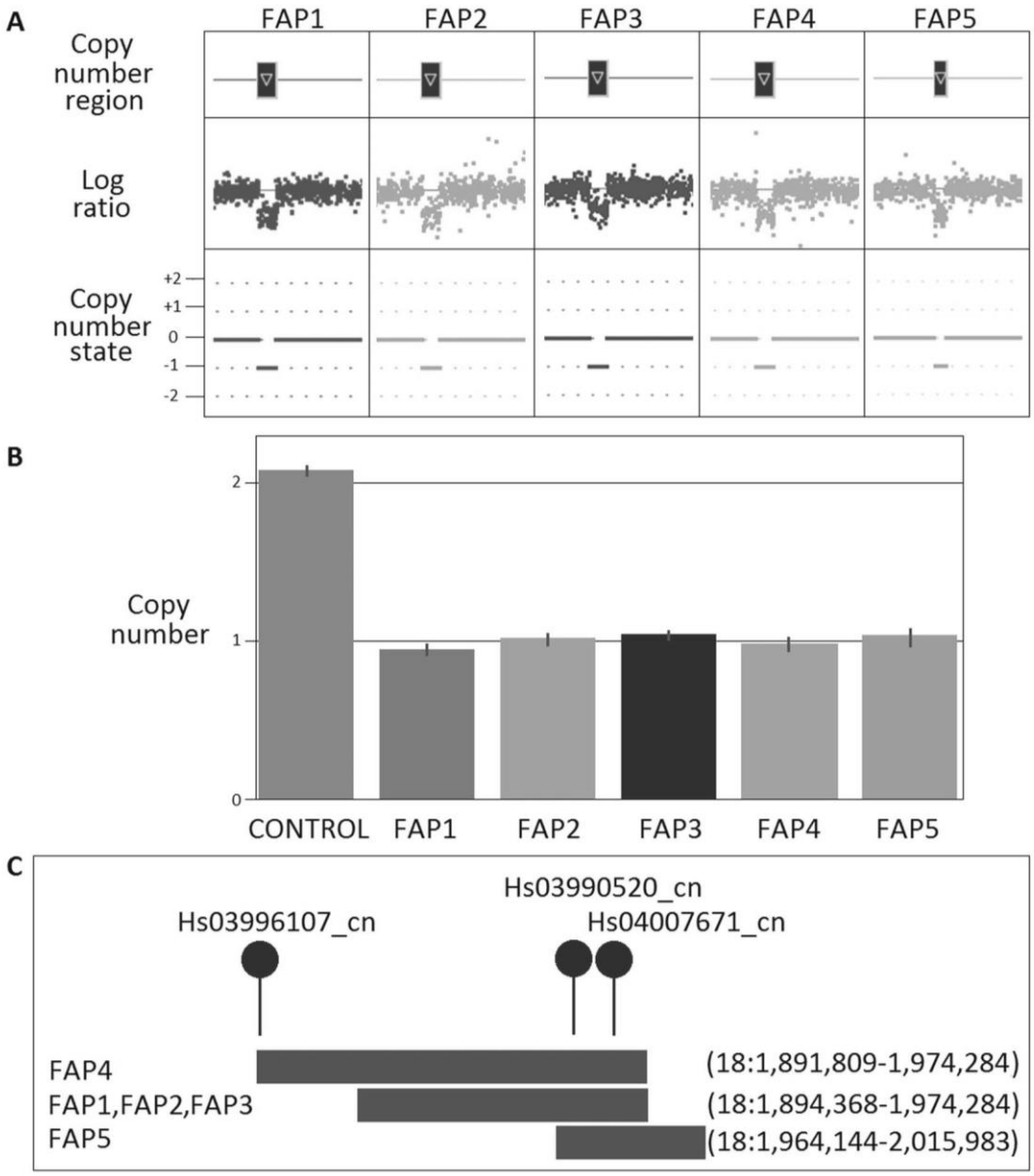


Figure 3 CNV results for 18p11.32 deletion in the FAP patients (FAP1, FAP2, FAP3, FAP4 and FAP5). (A) CNV profile from Cyto2.7M array data using ChAS noting the defined CN region (dark box above gene representing the CN deletion) in relation to the log ratio plot (relative fluorescence of each probe, dot, on the array showing a decrease in fluorescence indicating a loss in genomic material) and the CN state (0= normal two copies present, +1= one extra copy, +2= two extra copies, -1=one less copy and -2= two less copies); (B) Validation using TaqMan CN assay showing results for assay Hs03990520_cn noting the normal two copies of this region identified in the control (CON2), confirmation of the aberrant one copy in all affected FAP patients and the error bars associated with the three technical repeats for each sample; and (C) Location of CN deletions with respect to each other and the TaqMan CN assays used in validating the variants.

CRC susceptibility gene interrogation

We determined whether CNVs within 100 Kb either side of 77 genes involved in pathways known to be associated with CRC risk, including members of the WNT signalling and MMR pathways (see supplementary table 4 for a full list of genes), could potentially contribute to CRC development. Two unrelated polyposis patients harboured a CNV in the vicinity of the *MLH1* and *CTNNB1* genes, respectively. One patient harboured a CNV loss 18 Kb upstream *MLH1* located 3:36,925,248-36,991,856 (66.6 Kb, 95% confidence and detected by 35 probes) while the other harboured a CNV loss extending upstream and into the promoter region of *CTNNB1* located 3:41,068,578-41,119,502 (50.9 Kb, 94% confidence and detected by 24 probes). A third polyposis patient was identified harbouring a CNV loss located directly within the promoter 1B region of the *APC* tumour suppressor gene (5:112,065,033-112,096,002, 31 Kb; 91% confidence and detected by 56 probes) (figure 4). TaqMan CN assays confirmed the loss of this region in the affected patient as well as her two affected sons indicating this CN loss has been transmitted from one generation to the next. The loss of this region is likely to have contributed to disease development in all affected individuals.

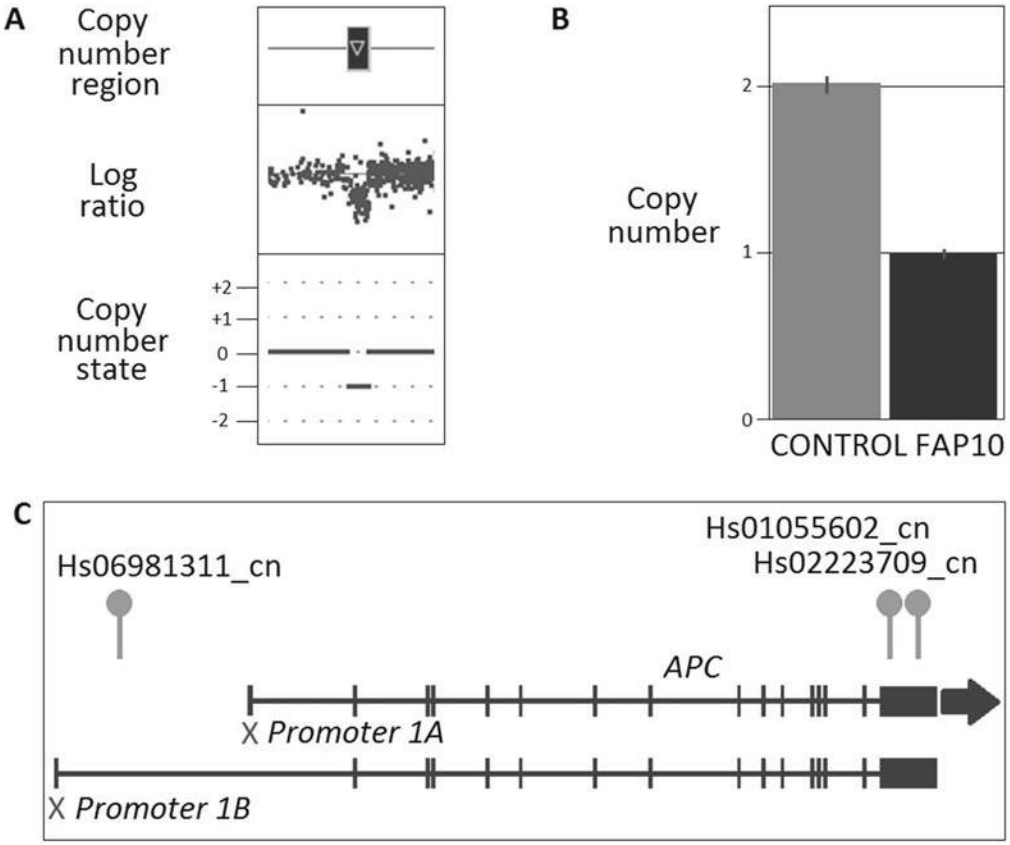


Figure 4 CNV results for the *APC* promoter 1B deletion in the FAP patient (FAP10). (A) CNV profile from Cyto2.7M array data using ChAS noting the defined CN region (dark box above gene representing the CN deletion) in relation to the log ratio plot (relative fluorescence of each probe, dot, on the array showing a decrease in fluorescence indicating a loss in genomic material) and the CN state (0= normal two copies present, +1= one extra copy, +2= two extra copies, -1=one less copy and -2= two less copies); (B) Validation using TaqMan CN assay showing results for assay Hs06981311_cn, noting the normal two copies of this region identified in the control (CON2), confirmation of the aberrant one copy in all affected FAP patient and the error bars associated with the three technical repeats for each sample; and (C) Location of CN duplication with respect to the gene and the TaqMan CN assays used in validating the variants.

Rare CNV events

The CNV dataset was also compared against the DGV. CNVs that were rare (not identified in the DGV and herein termed rare CNVs) corresponded to 31.29% (87 of 278) of the total CNVs identified in both patients and controls. In the control cohort 28.85% (30 of 104) of CNVs detected in 19 of the 40 controls (32.56%) were classified as rare whereas 32.76% (57 of 174) of CNVs detected in 23 of the 56 patients (41.07%) were rare. No significant difference was detected in the number of rare CNVs between patients and controls ($p=0.57$). In total, 49 genes were associated with the 57 rare CNVs identified in the polyposis cohort representing genes most likely to be associated with disease (see table 4). With the exception of *ANKFN1, FAM184B*, and *HCN1* (CNV loss) and *CCDC19* (CNV gain) which were not expressed in normal colon tissue and four other genes (*SNORD12, SNORD12B, SNORD12C, C20orf199*) where no information was available, 41 genes have been reported to be expressed in the colon and rectum (www.proteinatlas.org).

Table 4 List of 49 genes which may be implicated in disease (unique to the polyposis patients and not observed in the DGV). Note the gene symbol, description and if the gene is normally expressed in the colon (www.proteinatlas.org).

Gene	Description	Expression
ADD3	adducin 3 (gamma)	High
AIM1	absent in melanoma 1	Medium
AMICA1	adhesion molecule, interacts with CXADR antigen 1	Medium
ANKFN1	ankyrin-repeat and fibronectin type III domain containing 1	Not Detected
APC	adenomatous polyposis coli	Low
ARHGAP25	Rho GTPase activating protein 25	Low
ARHGAP26	Rho GTPase activating protein 26	Medium
ARHGDIB	Rho GDP dissociation inhibitor (GDI) beta	High
BCL2A1	BCL2-related protein A1	Low
BOD1L	biorientation of chromosomes in cell division 1-like 1	Medium
C17orf95	methyltransferase like 23	Medium
C20orf199	ZNFX1 antisense RNA 1	Unknown
C5orf56	chromosome 5 open reading frame 56	Low
CCDC19	coiled-coil domain containing 19	Not Detected
CDH11	cadherin 11, type 2, OB-cadherin (osteoblast)	Medium
CEACAM6	carcinoembryonic antigen-related cell adhesion molecule 6 (non-specific cross reacting antigen)	High
DDX10	DEAD (Asp-Glu-Ala-Asp) box polypeptide 10	Medium
ETV6	ets variant 6	Medium
FAM184B	family with sequence similarity 184, member B	Not Detected
FKBP1A	FK506 binding protein 1A, 12kDa	High
G0S2	G0/G1switch 2	Medium
GALC	galactosylceramidase	Medium
GAS7	growth arrest-specific 7	Medium
GPR65	G protein-coupled receptor 65	Low
HCN1	hyperpolarization activated cyclic nucleotide-gated potassium channel 1	Not Detected
HSD11B1	hydroxysteroid (11-beta) dehydrogenase 1	Low

JMJD6	jumonji domain containing 6	Medium
LAMB3	laminin, beta 3	Medium
LCA5	Leber congenital amaurosis 5	Low
MFSD11	major facilitator superfamily domain containing 11	Medium
MPZL3	myelin protein zero-like 3	Low
MXRA7	matrix-remodelling associated 7	Medium
NAMPT	nicotinamide phosphoribosyltransferase	Medium
NUMA1	nuclear mitotic apparatus protein 1	High
PLEK	pleckstrin	Medium
QKI	QKI, KH domain containing, RNA binding	Medium
RFT1	RFT1 homolog (S. cerevisiae)	Medium
SELL	selectin L	Low
SFRS2	arginine/serine rich splicing factor 2	High
SH3BGRL2	SH3 domain binding glutamic acid-rich protein like 2	High
SMAP2	small ArfGAP2	Medium
SMC3	structural maintenance of chromosomes 3	Medium
SNORD12	small nucleolar RNA, C/D box 12	Unknown
SNORD12B	small nucleolar RNA, C/D box 12B	Unknown
SNORD12C	small nucleolar RNA, C/D box 12C	Unknown
STK17B	serine/threonine kinase 17b	Medium
STX8	syntaxin 8	High
TAGLN2	transgelin 2	High
ZNFX1	zinc finger, NFX1-type containing 1	Medium

Pathway analysis and miR annotation of rare genes in polyposis patients

The 45 of the 49 rare genes were successfully mapped and investigated further using WebGestalt pathway analysis[19] for enrichment among KEGG pathways, cytogentic band regions and 3'UTR regions of genes (i.e. miR targets).

KEGG analysis revealed no significant pathways however, enrichment among cytogenetic bands identified 11 cytobands (20q13, 17q25, 6q14, 20q, 10q25, 6q, 2p13, 11q23, 1q32, 17q and 5q31; all with $p<0.0443$; see supplementary table 5) that were associated with 22 of the 49 rare genes (see supplementary table 6). Enrichment analysis for the targets of miRs identified 26 significant regions (all with $p<0.0421$) within the 3'UTR of 19 of the 49 rare genes for which 42 miRs were suggested to target (supplementary tables 5 and 6).

TAM[20] of the 42 miRs subsequently identified a total of 161 miR categories: 10 families, 13 clusters, 34 functional categories, 102 Human miR associated disease database (HMDD) and 2 tissue specificity categories. Specifically, miRs were significantly over-represented in the family category miR-15 (miR-15a, miR-195, miR-15b and miR-16; $p=0.002$) and the HMDD categories for glioblastoma (miR-15a, miR-195, miR-16, miR-181a-c, miR-18a and miR-32; $p=0.0205$), breast neoplasia (miR-15a, miR-520a, miR-320, miR-200a, miR-181b, miR-135b, miR-135b, miR-497, miR-133a, miR-27a, miR-302c, miR18a, miR-195, miR-30a, miR-18b and miR-519c; $p=0.002$) and leukaemia (mir-15a, miR-181a, miR-181b and miR-16; $p=0.008$).

Discussion

CNVs have yet to be intensively investigated for their involvement in polyposis and consequent contribution to disease development. Here we have presented a comparison between 56 unrelated *APC* and *MUTYH* mutation negative patients all diagnosed with polyposis and 40 healthy controls. We have furthermore compared our results the COSMIC database and DGV and have assessed our data in terms of CNV abundance, size and distribution using a whole genome approach in search of genes and genomic regions that could be associated with polyposis.

An increased CNV burden has been suggested to be associated with an increased risk of disease development, while variation in CNV burden is associated with phenotypic variation[22]. We did not identify any significant differences in the number or size of CNVs between polyposis patients and controls. This finding suggests that the numerical burden of CNVs (>6.03 Kb) does not appear to contribute to an increased disease risk. However, since our analysis was limited to the detection of CNVs greater than 6.03 Kb we cannot rule out the involvement of smaller CNVs in the aetiology of this disease.

This study revealed several CNVs affecting recurrent loci that included genes known to be associated with CRC, in multiple patients: *DOCK4* variants have been reported to give rise to various cancers including ovarian, prostate, glioma and CRC[23]; *DOCK4* is also involved with the regulation of □-catenin in the WNT signalling pathway[24], which has been directly implicated in FAP development; *NAMPT*, which is involved in the metabolism and proliferation of cells[25], has been identified to be over-expressed in CRC[26]; *NAMPT* is also a target of mir-26b (a putative tumour suppressor-miR) which binds to the 3' UTR of *NAMPT*[27]; while duplications in the *NF1* gene have also been observed in CRCs[28] suggesting that CNVs associated with these genes contribute to disease.

Furthermore, two of the recurrent CNV deletions observed in the current study also fell into regions containing recurrent deletions peaks reported by the TCGA for colon adenocarcinoma tumour data (located 16p13.3 and 7q 31.3), one of which is reported to harbour the disease candidate gene *A2BP1*[29]. Several other CNV regions identified among patients in the present study were also recurrent in the TCGA dataset (located 11q22.3, 15q21.1, 1p33, 20p12.1, 5q22.2, 7q31.3 and 5p12) containing several candidate disease genes including *APC*, *B2M*, *AGBL4*, *MACROD2* and *HCN1*[29]. Overall the results from the current study are consistent with previous reports on CNV

burden in CRC, however our data suggests several additional genomic regions may contribute to disease in these polyposis patients.

The distribution of CNVs on individual chromosomes was also compared between patients and controls, which failed to reveal any significant difference in the frequency of total CNVs between the two groups. The frequency of CNVs on each chromosome could not be shown to be significantly different, but a trend was observed indicating a greater number of CNVs on chromosome 18 compared to the controls. Among these CNVs was a gain in the *DCC* gene (identified in one patient). *DCC* is reported as a tumour suppressor and is frequently observed to be down-regulated in CRC (~70% of patients) which has been attributed to the loss of genomic material in the 18q21 region in which *DCC* resides[30,31]. Here we report the possible loss of *DCC* gene function as a result of an intronic CN gain. It has been revealed by others that some deep intronic variants have been shown to contribute to CRC via the formation of pseudoexons, the activation of cryptic splice sites and the expression of aberrant mRNA transcripts[32]. Validation studies using TaqMan CN assays confirmed the CN gain in the affected patient, however further studies will be required to understand the role of *DCC* in CRC.

Intriguingly CN losses rather than CN gains were shown to be statistically enriched on chromosome 18 in polyposis patients. We observed a region of CN loss at 18p11.32 that affects nearly 9% of the polyposis patients in our study. GWAS and meta-analysis studies have previously recognized 18p11.32 as susceptibility loci for bipolar disease, childhood acute lymphoblastic leukemia (ALL) and leisure time exercise behaviour[33-35]. More recently, loss of heterozygosity (LOH) at 18p11.32 has been reported in CRC adenomas (but not normal mucosa) and is suggested to be involved in CRC tumour development[36]; and a second study reporting genomic losses at 18p11.32 in CRCs are suggesting that this region is associated with adenoma-carcinoma progression[37]. Of particular note was the occurrence of the recently reported lncRNA (TCONS_00026231) residing in this region of loss. LncRNAs (non-coding nucleotides, 200-100,000 bp in size) are proposed to be master regulators whose functions include post-transcriptional regulation of gene expression, regulation of epigenetic marks, gene activation in *cis* and they have been shown to influence processes such as pluripotency[38-40]. Validation studies using TaqMan CN assays confirmed the CN losses in all affected patients. The frequency of this variant in a series of polyposis patients suggests that it may be associated with a predisposition of CRC in a proportion of *APC/MUTYH* mutation negative patients. Further studies are required to ascertain the precise involvement of this lncRNA in the genesis of CRC and more

specifically whether it is involved in controlling WNT signalling and therefore adenomatous polyposis development.

Investigation of CNVs residing in or in the proximity of known cancer genes or pathways may expand our understanding of their contribution to disease risk in polyposis. Herein we interrogated the CNV data in search for variants associated with genes in the WNT signalling and MMR pathways focusing on *APC* and *MUTYH*[6,18]. CNVs arising in or in the proximity of any of these genes may contribute to disease directly or via more cryptic means. Two unrelated polyposis patients harboured CNVs near *MLH1* and *CTNNB1.* Germline variants arising in *MLH1* are typically associated with Lynch syndrome[3]; whereas mutations associated with *CTNNB1* occur in sporadic CRC and other malignancies[41,42]; furthermore mutations in *CTNNB1* are reported to be enriched in desmoid disease[43], the second major cause of mortality in FAP. Interestingly the TCGA results on colon adenocarcinoma also reports *CTNNB1* as one of the most significantly mutated genes in (5%) non-hypermutated colon tumours[29]. The involvement of these two genes is concordant with both Lynch syndrome and FAP, respectively[44].

We identified a genomic loss located directly within the promoter 1B region of *APC*. It has been estimated that up to 2% of the mutations identified in FAP cases are large deletions, including deletions that extend from the promoter into the coding region[45]. Of *APC*s two promoter regions, promoter 1A and 1B, the latter of these is suggested to only play a minor role in *APC* gene regulation[46]. The first report attempting to characterize promoter-specific deletions in FAP was described in 2008[47], while Rohlin *et al*.[48] has recently reported the first evidence of promoter 1B involvement in FAP which was associated with a partial deletion of this region. Validation studies using TaqMan CN assays confirmed the CN loss in the affected patient. As the patient's two affected sons were also verified to carry the same CN loss confirming that the variant was transmitted from one generation to the next, this CN is likely to be the cause of disease in all the affected family members. Our study further supports the role of *APC* promoter 1B inactivation in FAP development, which is reinforced by the finding that the CNV is transmitted across generations and is suggested to segregate with the expected phenotype. It should also be noted that deep intronic mutations (smaller than the level of detection) in the *APC* gene and low level somatic mosaicism are reported to account for a proportion of polyposis patients[32,49] and may remain a possible unexplored cause of disease in a fraction of the other patients in this cohort.

In this study we also compared CNV data from the patient cohort to the DGV, a much larger control database than that available from the current study. A list of 49 rare genes was revealed that were likely to be associated with disease. *In silico* analysis was undertaken in search for biologically meaningful relationships among these 49 genes to provide insight into their potential contribution to disease. Several cytogenetic bands that were enriched in the analysis have previously been associated with CRC (20q13, 20q, 10q25, 6q, 11q23, 1q32 and 5q31)[50-56]. Of particular interest was the enrichment of 11q23 which has recently been reported to harbour risk variants for genetically unexplained colorectal adenomatous polyposis[57]. The findings in the current study provide further support for the possible involvement of this region disease.

Annotation of the 42 miRs proposed to target the enriched 3'UTR miR target region of the 49 rare genes further identified the miR-15 family to be overrepresented which is particularly interesting given this family is has been reported to be associated with tumour suppression[58,59]. In CRC more specifically, targeting the miRs miR-15 and miR-16 has been suggested as an effective mechanism to inhibit the growth of CRCs[60].

In conclusion, this study has revealed a number of CNVs which may contribute to the identification of genes and genomic regions associated with polyposis development and/or progression. Microarray analysis has identified several previously reported CRC susceptibility genes affected by CNVs in several patients, including *MLH1*, *CTNNB1* and *APC*. We have also identified chromosome 18 to be a region of interest since loss of 18p11.32 in multiple unrelated patients is associated with a lncRNA that may be involved in disease development. Overall the results of this study provide further evidence for the involvement of CNVs in the aetiology of polyposis.

Author contributions

ALM conducted the experiments, performed data analysis/interpretation and wrote the first draft of the manuscript.

BAT-P and T-JE provided expertise in data analysis and interpretation as well as revising the manuscript.

PM provided statistical expertise.

ADS enabled patient recruitment into this study.

GNH provided critical review of the manuscript and helped design the experiments.

RJS conceived the study, designed the experimental approach and reviewed and approved the final version of the manuscript prior to submission.

Additional Information

This work has been supported by the following funding bodies and Institutions: Australian Rotary Health/Rotary District 9650, the Commonwealth Scientific and Industrial Research Organization (CSIRO), the University of Newcastle and the Hunter Medical Research Institute.

Competing Financial Interests

The author(s) declare no competing financial interests.

List of abbreviations

ALL	acute lymphoblastic leukaemia
BH	Bengamini and Hochberg
CHAS	Chromosome Analysis Suite
CN	copy number
CNV	copy number variation
COSMIC	Catalogue of Somatic Mutations in Cancer
CRC	colorectal cancer
DGV	Database of genomic variants
DNA	deoxyribose nucleic acid
FAP	familial adenomatous polyposis
HMDD	human microRNA disease database
Kb	kilobase
KEGG	Kyoto Encyclopaedia of Genes and Genomes
lncRNA	link RNA
LOH	loss of heterozygosity
mapd	median absolute pairwise difference
miR	microRNA
MLPA	multiplex ligation-dependant probe amplification
MMR	mismatch repair
ng	nanogram
NTC	no template control
QC	quality control
RNA	ribose nucleic acid

SNP	single nucleotide polymorphism
TAM	Tool for the annotation of microRNAs
TCGA	The Cancer Genome Atlas
UCSC	University of California, Santa Cruz

References

1. Galiatsatos, P. & Foulkes, W.D. Familial adenomatous polyposis. *Am J Gastroenterol* **101**, 385-98 (2006).

2. Lynch, H.T. & de la Chapelle, A. Hereditary colorectal cancer. *N Engl J Med* **348**, 919-32 (2003).

3. Rustgi, A.K. The genetics of hereditary colon cancer. *Genes Dev* **21**, 2525-38 (2007).

4. Fokkema, I.F. *et al.* LOVD v.2.0: the next generation in gene variant databases. *Hum Mutat* **32**, 557-63 (2011).

5. Stenson, P.D. *et al.* The Human Gene Mutation Database: 2008 update. *Genome Med* **1**, 13 (2009).

6. Pezzi, A. *et al.* Relative role of APC and MUTYH mutations in the pathogenesis of familial adenomatous polyposis. *Scand J Gastroenterol* **44**, 1092-100 (2009).

7. Stella, A. *et al.* Germline novel MSH2 deletions and a founder MSH2 deletion associated with anticipation effects in HNPCC. *Clin Genet* **71**, 130-9 (2007).

8. Morak, M. *et al.* Biallelic MLH1 SNP cDNA expression or constitutional promoter methylation can hide genomic rearrangements causing Lynch syndrome. *J Med Genet* **48**, 513-519 (2011).

9. Clendenning, M. *et al.* Mutation deep within an intron of MSH2 causes Lynch syndrome. *Fam Cancer* **10**, 297-301 (2011).

10. Chan, T.L. *et al.* A novel germline 1.8-kb deletion of hMLH1 mimicking alternative splicing: a founder mutation in the Chinese population. *Oncogene* **20**, 2976-81 (2001).

11. Ligtenberg, M.J. *et al.* 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 (2009).

12. Hochstenbach, R. *et al.* Discovery of variants unmasked by hemizygous deletions. *Eur J Hum Genet* **20**, 748-53 (2012).

13. Shlien, A. & Malkin, D. Copy number variations and cancer susceptibility. *Curr Opin Oncol* **22**, 55-63 (2010).

14. Krepischi, A.C. *et al.* Germline DNA copy number variation in familial and early-onset breast cancer. *Breast Cancer Res* **14**, R24 (2012).

15. Pylkas, K. *et al.* Rare copy number variants observed in hereditary breast cancer cases disrupt genes in estrogen signaling and TP53 tumor suppression network. *PLoS Genet* **8**, e1002734 (2012).

16. McEvoy, M. *et al.* Cohort profile: The Hunter Community Study. *Int J Epidemiol* **39**, 1452-63 (2010).

17. Miller, S.A., Dykes, D.D. & Polesky, H.F. A simple salting out procedure for extracting DNA from human nucleated cells. *Nucleic Acids Res* **16**, 1215 (1988).

18. Molatore, S. *et al.* MUTYH mutations associated with familial adenomatous polyposis: functional characterization by a mammalian cell-based assay. *Hum Mutat* **31**, 159-66 (2010).

19. Zhang, B., Kirov, S. & Snoddy, J. WebGestalt: an integrated system for exploring gene sets in various biological contexts. *Nucleic Acids Res* **33**, W741-8 (2005).

20. Lu, M., Shi, B., Wang, J., Cao, Q. & Cui, Q. TAM: a method for enrichment and depletion analysis of a microRNA category in a list of microRNAs. *BMC Bioinformatics* **11**, 419 (2010).

21. Hetzenecker, A.M. *et al.* Downregulation of the ubiquitin-proteasome system in normal colonic macrophages and reinduction in inflammatory bowel disease. *Digestion* **86**, 34-47 (2012).

22. Girirajan, S. & Eichler, E.E. Phenotypic variability and genetic susceptibility to genomic disorders. *Hum Mol Genet* **19**, R176-87 (2010).

23. Kuo, K.T. *et al.* Analysis of DNA copy number alterations in ovarian serous tumors identifies new molecular genetic changes in low-grade and high-grade carcinomas. *Cancer Res* **69**, 4036-42 (2009).

24. Upadhyay, G. *et al.* Molecular association between beta-catenin degradation complex and Rac guanine exchange factor DOCK4 is essential for Wnt/beta-catenin signaling. *Oncogene* **27**, 5845-55 (2008).

25. Zhang, L.Y. *et al.* Anti-proliferation effect of APO866 on C6 glioblastoma cells by inhibiting nicotinamide phosphoribosyltransferase. *Eur J Pharmacol* **674**, 163-70 (2012).

26. Hufton, S.E. *et al.* A profile of differentially expressed genes in primary colorectal cancer using suppression subtractive hybridization. *FEBS Lett* **463**, 77-82 (1999).

27. Zhang, C., Tong, J. & Huang, G. Nicotinamide phosphoribosyl transferase (Nampt) is a target of microRNA-26b in colorectal cancer cells. *PLoS One* **8**, e69963 (2013).

28. Cacev, T., Radosevic, S., Spaventi, R., Pavelic, K. & Kapitanovic, S. NF1 gene loss of heterozygosity and expression analysis in sporadic colon cancer. *Gut* **54**, 1129-35 (2005).

29. Cancer Genome Atlas, N. Comprehensive molecular characterization of human colon and rectal cancer. *Nature* **487**, 330-7 (2012).

30. Fearon, E.R. *et al.* Identification of a chromosome 18q gene that is altered in colorectal cancers. *Science* **247**, 49-56 (1990).

31. Thiagalingam, S. *et al.* Evaluation of candidate tumour suppressor genes on chromosome 18 in colorectal cancers. *Nat Genet* **13**, 343-6 (1996).

32. Spier, I. *et al.* Deep intronic APC mutations explain a substantial proportion of patients with familial or early-onset adenomatous polyposis. *Hum Mutat* **33**, 1045-50 (2012).

33. De Moor, M.H. *et al.* Genome-wide association study of exercise behavior in Dutch and American adults. *Med Sci Sports Exerc* **41**, 1887-95 (2009).

34. Ferreira, M.A. *et al.* Collaborative genome-wide association analysis supports a role for ANK3 and CACNA1C in bipolar disorder. *Nat Genet* **40**, 1056-8 (2008).

35. Trevino, L.R. *et al.* Germline genomic variants associated with childhood acute lymphoblastic leukemia. *Nat Genet* **41**, 1001-5 (2009).

36. Costi, R. *et al.* Repeated anastomotic recurrence of colorectal tumors: genetic analysis of two cases. *World J Gastroenterol* **17**, 3752-8 (2011).

37. Shi, Z.Z. *et al.* Genomic profiling of rectal adenoma and carcinoma by array-based comparative genomic hybridization. *BMC Med Genomics* **5**, 52 (2012).

38. Loewer, S. *et al.* Large intergenic non-coding RNA-RoR modulates reprogramming of human induced pluripotent stem cells. *Nat Genet* **42**, 1113-7 (2010).

39. Nagano, T. *et al.* The Air noncoding RNA epigenetically silences transcription by targeting G9a to chromatin. *Science* **322**, 1717-20 (2008).

40. Orom, U.A. *et al.* Long noncoding RNAs with enhancer-like function in human cells. *Cell* **143**, 46-58 (2010).

41. Hirata, H. *et al.* MicroRNA-1826 targets VEGFC, beta-catenin (CTNNB1) and MEK1 (MAP2K1) in human bladder cancer. *Carcinogenesis* **33**, 41-8 (2012).

42. Sygut, A. *et al.* Genetic Variations of the CTNNA1 And The CTNNB1 Genes in Sporadic Colorectal Cancer in Polish Population. *Pol Przegl Chir* **84**, 560-4 (2012).

43. Le Guellec, S. *et al.* CTNNB1 mutation analysis is a useful tool for the diagnosis of desmoid tumors: a study of 260 desmoid tumors and 191 potential morphologic mimics. *Mod Pathol* **25**, 1551-8 (2012).

44. Jasperson, K.W., Tuohy, T.M., Neklason, D.W. & Burt, R.W. Hereditary and familial colon cancer. *Gastroenterology* **138**, 2044-58 (2010).

45. Gismondi, V. *et al.* 310 basepair APC deletion with duplication of breakpoint (439ins15del310) in an Italian polyposis patient. *Hum Mutat* **Suppl 1**, S220-2 (1998).

46. Tsuchiya, T. *et al.* Distinct methylation patterns of two APC gene promoters in normal and cancerous gastric epithelia. *Oncogene* **19**, 3642-6 (2000).

47. Charames, G.S. *et al.* A large novel deletion in the APC promoter region causes gene silencing and leads to classical familial adenomatous polyposis in a Manitoba Mennonite kindred. *Hum Genet* **124**, 535-41 (2008).

48. Rohlin, A. *et al.* Inactivation of promoter 1B of APC causes partial gene silencing: evidence for a significant role of the promoter in regulation and causative of familial adenomatous polyposis. *Oncogene* (2011).

49. Aretz, S. *et al.* Somatic APC mosaicism: a frequent cause of familial adenomatous polyposis (FAP). *Hum Mutat* **28**, 985-92 (2007).

50. Houlston, R.S. *et al.* 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 (2010).

51. Jia, W.H. *et al.* Genome-wide association analyses in East Asians identify new susceptibility loci for colorectal cancer. *Nat Genet* **45**, 191-6 (2013).

52. Jiao, S. *et al.* Genome-wide search for gene-gene interactions in colorectal cancer. *PLoS One* **7**, e52535 (2012).

53. Peters, U. *et al.* Meta-analysis of new genome-wide association studies of colorectal cancer risk. *Hum Genet* **131**, 217-34 (2012).

54. Peters, U. *et al.* Identification of Genetic Susceptibility Loci for Colorectal Tumors in a Genome-Wide Meta-analysis. *Gastroenterology* **144**, 799-807 e24 (2013).

55. Tenesa, A. *et al.* 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 (2008).

56. Cui, R. *et al.* Common variant in 6q26-q27 is associated with distal colon cancer in an Asian population. *Gut* **60**, 799-805 (2011).

57. Hes, F.J. *et al.* Colorectal cancer risk variants on 11q23 and 15q13 are associated with unexplained adenomatous polyposis. *J Med Genet* **51**, 55-60 (2014).

58. Calin, G.A. *et al.* Frequent deletions and down-regulation of micro- RNA genes miR15 and miR16 at 13q14 in chronic lymphocytic leukemia. *Proc Natl Acad Sci U S A* **99**, 15524-9 (2002).

59. Roccaro, A.M. *et al.* MicroRNAs 15a and 16 regulate tumor proliferation in multiple myeloma. *Blood* **113**, 6669-80 (2009).

60. Dai, L. *et al.* Vector-based miR-15a/16-1 plasmid inhibits colon cancer growth in vivo. *Cell Biol Int* **36**, 765-70 (2012).

CHAPTER 3: COPY NUMBER VARIATION IN HNPCC

Introduction

Hereditary non-polyposis colorectal cancer (HNPCC) represents somewhere between 2% and 5% of all colorectal cancers and is attributed to heritable germline alterations that result in the inactivation of one of four DNA mismatch repair (MMR) genes: *MLH1*, *MSH2*, *MSH6* or *PMS2*[63-69]. Up to 50% of clinically tested patients are found to harbor germline variants in these genes and are thereafter referred to as specifically having Lynch syndrome (LS)[128,129,183]. For the remaining 50% of patients no germline mutation can be identified, suggesting that either other genes or other mechanisms associated with the silencing of any of the four MMR genes could be responsible for LS.

Since the sequencing of the human genome it has become apparent that genomic rearrangements are ubiquitous in the population. Genomic duplication or deletion have been shown to encompass large stretches of contiguous DNA and are commonly termed copy number variants (CNVs). As CNVs range from kilobase (Kb) to megabase (Mb) in size, they may encompass and disrupt large amounts of DNA sequence that may result in altered gene expression and the development of disease[278,343-348].

Recent reports specifically examining the association between genomic rearrangements and LS have revealed a loss on chromosome 2 encompassing the polyadenylation signal of *EPCAM* which results in the transcriptional silencing of *MSH2* via transcription read-though[144,349]. This evidence suggests that a proportion of HNPCC families may be accounted for by genomic rearrangements that may not be identified by the highly targeted genetic screening used in a clinical setting.

This part of the thesis aims to describe the CNV landscape in patients without identifiable mutations in any of the four MMR genes commonly associated with LS. Furthermore genes associated with unique CNVs in patients will also be investigated using pathway analysis and miR annotation in an attempt to reveal biologically meaningful relationships which may underpin the development of disease in these patients.

Publication

Amy L. Masson, Bente A. Talseth-Palmer, Tiffany-Jane Evans, Desma M. Grice, Konsta Duesing, Garry N. Hannan and Rodney J. Scott (2013) Copy Number Variation in Hereditary Non-Polyposis Colorectal Cancer, *Genes*, *4*, 536-555.

Co-author statement

I attest that Research Higher Degree candidate Amy Louise Masson contributed to the above manuscript including involvement in the conception and design of the study; conducting the laboratory work, data analysis and interpretation; and preparation of the manuscript.

Co-author	Signature	Date
Bente A. Talseth-Palmer		
Tiffany-Jane Evans		
Desma M. Grice		
Konsta Duesing		
Garry N. Hannan		
Rodney J. Scott		

Amy Louise Masson Date: 01/09/2015

Professor Robert Callister Date: 01/09/2015

Assistant Dean Research Training

Copy Number Variation in Hereditary Non-Polyposis Colorectal Cancer

Amy L. Masson[1,4], Bente A. Talseth-Palmer[1,4], Tiffany-Jane Evans[1,4], Desma M. Grice[1,2], Konsta Duesing[2], Garry N. Hannan[2] and Rodney J. Scott[1,3,4,*]

[1] Information Based Medicine Program, Hunter Medical Research Institute, University of Newcastle, Newcastle, New South Wales, 2305 Australia; E-mails: c3027685@uon.edu.au (A.M.); Bente.Talseth-Palmer@newcastle.edu.au (B.TP.); Tiffany-Jane.Evans@newcastle.edu.au (TJ.E.); Desma.Grice@csiro.au (D.G.)

[2] Preventative Health Flagship and Animal, Food and Health Sciences Divisions, CSIRO, Ryde, New South Wales, 2113 Australia; E-mails: Konsta.Duesing@csiro.au (K.D.); Garry.Hannan@csiro.au (G.H.)

[3] Division of Molecular Medicine, Hunter Area Pathology Service, John Hunter Hospital, Newcastle, New South Wales, 2305 Australia

[4] School of Biomedical Sciences and Pharmacy, Faculty of Health, University of Newcastle, New South Wales, 2308 Australia

*Author to whom correspondence should be addressed; E-mail: Rodney.Scott@newcastle.edu.au (R.J.S.)

Tel.: +61 (2) 4921 4974

Fax: +61 (2) 4921 4253

Abstract

Hereditary non-polyposis colorectal cancer (HNPCC) is the commonest form of inherited colorectal cancer (CRC) predisposition and by definition describes families which conform to the Amsterdam Criteria or reiterations thereof. In ~50% of patients adhering to the Amsterdam criteria germline variants are identified in one of four DNA Mismatch repair (MMR) genes MLH1, MSH2, MSH6 and PMS2. Loss of function of any one of these genes results in a failure to repair DNA errors occurring during replication which can be most easily observed as DNA microsatellite instability (MSI) – a hallmark feature of this disease. The remaining 50% of patients without a genetic diagnosis of disease may harbour more cryptic changes within or adjacent to MLH1, MSH2, MSH6 or PMS2 or elsewhere in the genome. We used a high density cytogenetic array to screen for deletions or duplications in a series of patients, all of whom adhered to the Amsterdam/Bethesda criteria, to determine if genomic re-arrangements could account for a proportion of patients that had been shown not to harbour causative mutations as assessed by standard diagnostic techniques. The study has revealed some associations between CNVs and HNPCC mutation negative cases and further highlights difficulties associated with CNV analysis.

Keywords

Microsatellite instability (MSI); Cancer; DNA Repair; Diagnostic Testing; HNPCC/Lynch Syndrome; Copy Number Variation; Affymetrix; Array.

Introduction

Somewhere between 2% and 5% of all colorectal cancers (CRCs) are classified as hereditary non-polyposis colorectal cancer (HNPCC). Families with germline mutations or complex genomic changes (without structural gene alterations) that render one of four DNA mismatch repair (MMR) genes ineffective compose a subset of HNPCC known as Lynch Syndrome (LS).

The clinical diagnosis of HNPCC is defined by any one of several reiterations of the Amsterdam Criteria, first established in 1990 to enable the identification of the genetic basis of the disease[1]. As such mutations in *MLH1*, *MSH2*, *MSH6* and *PMS2* have been identified to account for all LS families[2-4]. Recently, loss of *EPCAM*, has been associated with transcriptional silencing of *MSH2*, and rare epimutations in *MLH1* have also been implicated in LS[5,6].

Despite the definition of HNPCC up to 50% of clinically tested patients with tumours demonstrating microsatellite instability (MSI), the hallmark phenotype of HNPCC, will fail to have any germline mutation identified in any one of the four MMR genes responsible for LS[7-9]. This suggests that there are either other genes associated with this disorder or different mechanisms of gene silencing responsible for HNPCC.

Since the sequencing of the human genome it has become apparent that genomic rearrangements are ubiquitous in the population. Genomic duplication or deletion have been shown to encompass large stretches of contiguous DNA and are commonly termed copy number variants (CNVs). As CNVs range from kilobase (Kb) to megabase in size, they may encompass or disrupt functional DNA sequences, result in gene amplification or loss, or alter epigenetic patterning[10]. As such, CNVs have been well documented in their contribution to disease development and variation in disease phenotype[11-16].

CNVs have been implicated in the development of many forms of CRC, e.g. germline deletion of two genes, *PTEN* and *BMPR1A* have been identified to be the cause of Juvenile Polyposis (JP) in four unrelated children[17], while genomic deletions in the genes *SMAD4*, *BMPR1A* and *PTEN* result in JP[18] and furthermore, the Leiden Open Variation Database (LOVD) database lists nearly 3000 mutations in four MMR genes associated with HNPCC, of which many are gains and losses of genomic material[19]. Recent reports specifically examining the association between genomic rearrangements and LS have revealed that loss of a region on chromosome 2 encompassing *EPCAM* appears to be associated with LS[6,20]. The loss of *EPCAM*

appears to re-write the epigenetic programming in the region such that the *MSH2* becomes silenced as a result of CpG methylation of the 5'promoter region. This evidence suggests that a proportion of HNPCC families may be accounted for by genomic rearrangements that may not be readily identified using more traditional gene mutation searches.

CNVs are detected using DNA arrays that comprise a series of oligonucleotides that represent evenly distributed markers across the entire genome. As the number of oligonucleotide markers has increased from a few hundred thousand to over five million, CNV resolution has improved such that ever smaller rearrangements can be detected in a single experiment. In this study we have used the Affymetrix Cytogenetic Whole Genome 2.7M (Cyto2.7M) array which contains over 400,000 SNP probes and greater than 2.1 million CNV probes (average spacing 1395 base pairs) to examine the CNV landscape in HNPCC patients and search for CN gains or CN losses which may reside in or in the vicinity of the 22 genes associated with DNA MMR. We also investigated genes and gene expression regulatory elements (microRNAs or miRs) associated with CNVs unique to the HNPCC patients using pathway analysis to determine if they may contribute to disease development.

Experimental section

Samples

Genomic DNA samples for the current study were obtained from HNPCC patients who had given informed consent for their DNA to be used for studies into their disease and control DNA samples from the Hunter Community Study (HCS)[21]. DNA was extracted from whole blood by the salt precipitation method[22]. The study was approved by the University of Newcastle's Human Research Ethics Committee (HREC) and the Hunter New England Human Research Ethics Committee (HNEHREC).

A sample size of 125 HNPCC patients was used for the current study. All HNPCC patients were clinically diagnosed as per the Amsterdam Criteria II[1,23] or the Bethesda Guidelines[24]. All patients had been diagnosed with CRC and were the first individual (proband) of their family to seek genetic testing. The samples were referred for routine clinical diagnostic testing involving screening for mutations in: *MLH1*, *MSH2*, *MSH6* and/or *PMS2*. The mutation screening was performed using Sanger Sequencing and/or Multiplex ligation-dependant probe amplification (MLPA). No mutations were identified in any of the patients used for the current study and are thus considered to be MMR mutation negative. The average age of patients recruited for this study was 52 years of age.

A sample size of 40 controls from the Hunter Community Study (HCS)[21] was used in the current study. Theses samples were from healthy individuals aged >55 years who were cancer free at the time of sample collection.

Genomic array analysis

The DNA from the 165 patients and controls was processed on the Affymetrix Cyto2.7M array according to manufacturer's protocols. CEL files obtained from scanning the Cyto2.7M array were analysed in the proprietary software from Affymetrix, the Chromosome Analysis Suite (ChAS) (Version CytoB-N1.2.0.232; r4280) using NetAffx Build 30.2 (Hg18) annotation. Quality control parameters were optimized and validated using a training set of 20 randomly selected samples (patients and controls). Identified CNV regions within the training set were assessed according to CNV call confidence, probe count, size, wavinessSd and by visual inspection for distinction from normal CN state. In addition, data was visually inspected to identify regions with low density of markers including centromeric and telomeric regions (supplementary table 1) which were excluded from analysis across all samples. The resultant thresholds were applied to all samples. Most of the thresholds were more stringent than default

settings, aiming to minimize the number of false-positive CNVs being included in the analysis. Briefly, all samples were subject to a series of quality cut-off measures: snpQC >1.1 (assesses quality of SNP probes with respect distances between the distribution of alleles AA, AB and BB alleles and larger differences are associated with an increased ability to identify a genotype; default), mapdQC <0.27 (Median Absolute Pair-wise Difference; assesses quality of CN probes with respect to a reference model file; default) and wavinessSd <0.1 (measure of standard deviation in data waviness; the GC content across the genome correlates with average probe intensities i.e. high GC probes are brighter than low GC probes on average, which creates waves in the data). CNV regions within all samples were then filtered using a set of CNV calling parameters: >90% confidence, autosomes only and a minimum number of 24 probes.

Statistical, pathway and annotation analysis

Refined ChAS CNV counts and CNV size for the patients were compared to the controls using a two tailed un-paired t-test in Graphpad Prism (Version 6)[25]. Gene enrichment analysis was performed using WebGestalt analysis software (Version 2013)[26]. This software was used to assess gene lists derived from the refined CNV results obtained from ChAS according to Gene Ontology (GO) categories, Kyoto Encyclopaedia of Genes and Genomes (KEGG) pathways and miR targets. Analysis was performed using hypergeometric statistical method, Benjamini and Hochberg (BH) correction for multiple testing (both default settings) and a biological significance threshold of <0.05 with a minimum of two genes per category required to assess any enrichment. TAM (Tool for Annotations of miRs) (Version 2)[27] software was used to annotate miRs according to miR family, cluster, function, Human miR associated disease categories (HMDD) and tissue specificity. Annotations were performed using the following parameters: all miRs in the TAM database were used as a background; to identify meaningful categories we looked at miR over-representation in all categories and analysis was limited to at least one miR in a given category. Enrichment analysis for miRs categories was conducted using hypergeometric testing and *p* values were corrected according to Bonferroni correction for multiple testing.

Results and Discussion

Recent studies have reported CNV's as relevant contributors to human diversity and cancer susceptibility[28-30]. This study further defines the contribution of CNVs to disease risk in HNPCC.

Resolution

Refinement of ChAS thresholds resulted in the final analysis of CNVs ranging from a minimum of 8.4 Kb to a maximum of 2722.5 Kb in size (see figure 1 for examples). CN gains ranged from 8.4 Kb to 2722 Kb in patients and 14.8 Kb to 1076.2 Kb in controls while CN losses ranged from 17.8 Kb to 529.2 Kb in patients and 16.8 Kb to 1205.7 Kb in controls. As such we cannot rule out the potential involvement of CNVs below the level of detection of the Cyto2.7M array, in the aetiology of HNPCC.

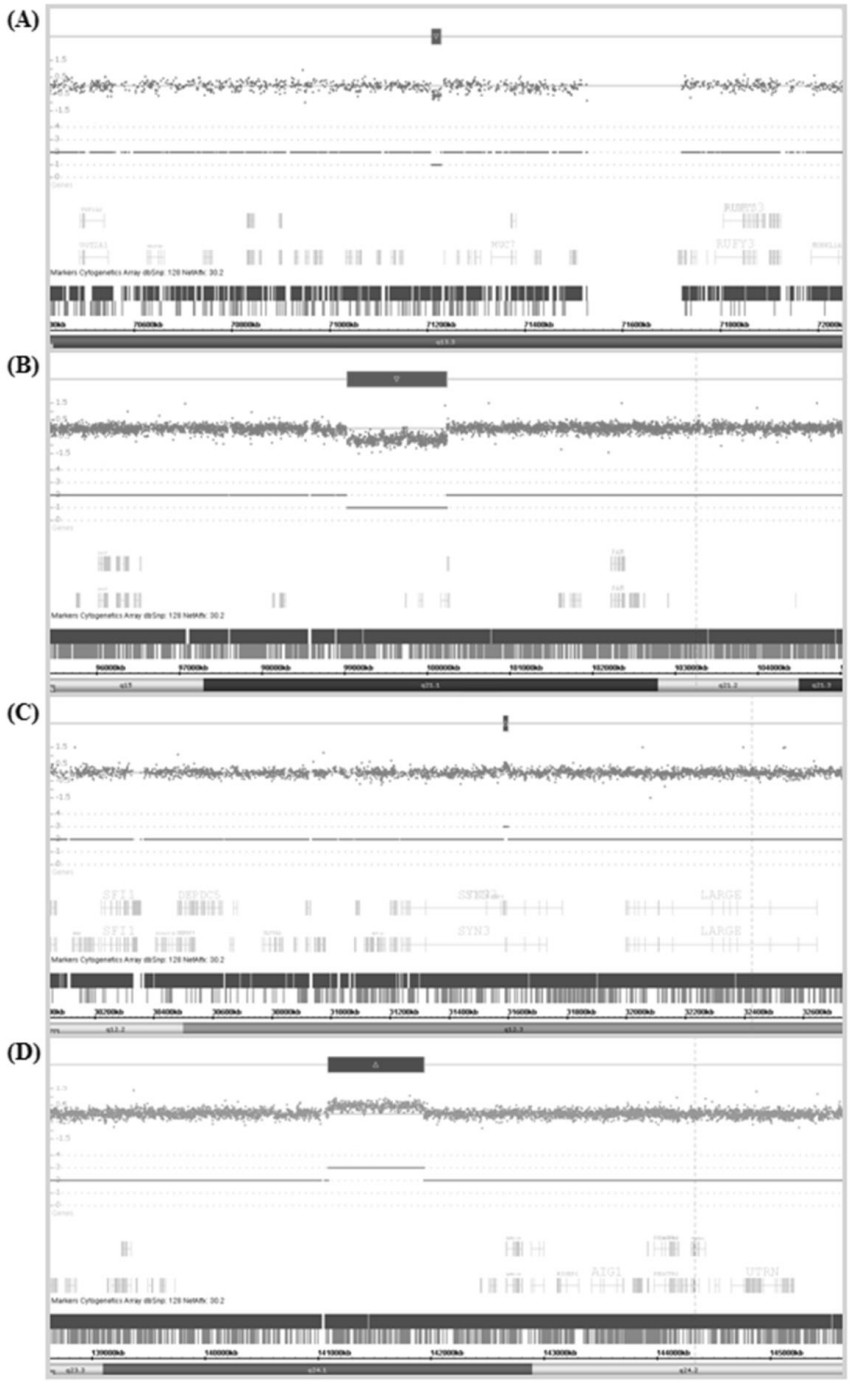


Figure 1 ChAS output showing examples of (A) a small CN loss of 17.8 Kb; (B) a large CN loss of 1205.7 Kb; (C) a small CN gain of 14.8 Kb; and (D) a large CN gain of 843.3 Kb. The Log2Ratio represents the relative fluorescence of each probe (dot) across the genome (from left to right). The fluorescence is reduced in regions of CN loss and increased in regions of CN gain. This is indicated by a CN loss or CN gain over the affected region and the resultant CN state noted below e.g. there is only one of the two alleles present in each CN loss and an extra allele present in each of the CN gains.

CNV Detection

Analysis of Cyto2.7M array data identified a total of 543 CNVs in the 165 patients and controls utilized in this study (table 1). Total counts of CNVs observed in the 125 HNPCC patients corresponded to 439 CNV events compared to 104 events in the 40 controls. The mean number of CNVs identified per sample did not significantly differ between patients and controls (3.51 CNVs per patient and 2.60 CNVs per control, *p=0.2980*). Consistent with a recent report looking at CNVs in hereditary breast cancer, similar counts of CNVs detected between patients and controls have been suggested to reflect a lack of genomic instability in the genomes of patients screened[31,32]. The mean CNV affected genome per sample did not differ between patients and controls either (284.07 Kb patients and 295.52 Kb controls, *p=0.9121*). However, the mean size of a CNV differed significantly between patients and controls (70.08 Kb in patients and 106.57 Kb in controls, *p=0.0165)*. The exact reason why we observe this difference is unclear however it may be a function of the number of samples in each group.

Table 1 Summary of CNV results obtained from the Cyto2.7M array analysed in ChAS.

		CNV Count			CNV Size (Kb)		
		Total CNVs per group	Median CNVs per sample	Mean CNVs per sample	Total CNV affected genome per group	Mean total CNV affected genome per sample	Mean size of a CNV
Patients	125	439	2	3.51	35,508.53	284.07	70.08
Controls	40	104	2	2.60	11,820.75	295.52	106.57
p	-	-	-	*0.2980*	-	*0.9121*	*0.0165**

*statistically significant

MMR gene interrogation

CNVs in patients and controls were interrogated for CN gains and losses residing in or in the vicinity of (50 Kb upstream to 50 Kb downstream) the 22 genes (*EXO1*, *LIG1*, *MLH1*, *MLH3*, *MSH2*, *MSH3*, *MSH6*, *PCNA, PMS1*, *PMS2*, *POLD1*, *POLD2*, *POLD3*, *POLD4*, *RFC1*, *RFC2*, *RFC3*, *RFC4*, *RFC5*, *RPA1*, *RPA2* and *RPA3*) in the MMR pathway (see table 2). We aimed to identify CNVs which could potentially contribute to disease development directly (e.g. disruption of functional gene sequences or promoter region inactivation) and via other mechanisms, including the alteration to epigenetic marks (as seen with the transcriptional silencing of *MSH2* through a CN loss in *EPCAM* in several LS patients, described previously[6,20]).

No CN gains or losses were identified within the defined search region for any of the 22 genes in the MMR pathway for all samples utilized in this study, patients and controls. We cannot however rule of the possibility for CNVs residing in these regions which are smaller than the resolution of detection provided by this array (<8.4 Kb).

Table 2 Regions searched for CN gains and CN losses in and in the vicinity of (±50 Kb) of the 22 genes in the MMR pathway. Chromosomal position of gene (start and end), gene size and search region (search start and search end) is noted.

Gene	Chr	Start (bp)	End (bp)	Size (Kb)	Search start (bp)	Search end (bp)
EXO1	1	240,078,157	240,119,671	42	240,028,157	240,169,671
RPA2	1	28,090,635	28,113,823	23	28,040,635	28,163,823
MSH2	2	47,783,766	47,563,864	80	47,733,766	47,613,864
MSH6	2	47,863,724	47,887,596	24	47,813,724	47,937,596
PMS1	2	190,357,055	190,450,600	94	190,307,055	190,500,600
MLH1	3	37,009,982	37,067,341	57	36,959,982	37,117,341
RFC4	3	187,990,375	188,007,178	17	187,940,375	188,057,178
RFC1	4	38,965,470	39,044,390	79	38,915,470	39,094,390
MSH3	5	79,986,049	80,208,390	222	79,936,049	80,258,390
PMS2	7	5,979,395	6,015,263	36	5,929,395	6,065,263
POLD2	7	44,120,810	44,129,672	9	44,070,810	44,179,672
RFC2	7	73,283,767	73,306,674	23	73,233,767	73,356,674
RPA3	7	7,643,099	7,724,763	82	7,593,099	7,774,763
POLD3	11	73,981,276	74,031,413	50	73,931,276	74,081,413
POLD4	11	66,875,594	66,877,593	2	66,825,594	66,927,593
RFC5	12	116,938,890	116,954,422	16	116,888,890	117,004,422
RFC3	13	33,290,205	33,438,695	148	33,240,205	33,488,695
MLH3	14	74,550,219	74,587,988	38	74,500,219	74,637,988
RPA1	17	1,680,022	1,749,598	70	1,630,022	1,799,598
LIG1	19	53,310,514	53,365,372	55	53,260,514	53,415,372
POLD1	19	55,579,404	55,613,083	34	55,529,404	55,663,083
PCNA	20	5,043,598	5,055,268	12	4,993,598	5,105,268

Occurrence and distribution of CNVs in patients and controls

Of the total 104 CNVs identified in controls, 34 CNVs contained genomic regions that were common to genomic regions identified in patients (supplementary table 2). A total of 70 CNVs were unique to the controls of which 47 (67.14%) were associated with genes (supplementary table 3).

Of 439 CNVs identified in patients, 53 CNVs contained genomic regions that were common to genomic regions identified in controls (supplementary table 4). 386 CNVs were unique to the patients population of which 207 (53.63%) were associated with genes (supplementary table 5).

From the 207 unique CNVs associated with genes identified in the patients, 9 were identified in patients that did not overlap any CNVs in controls but affected the same gene even in multiple patients (*ARPP-21*, *C7orf10*, *KIAA1217*, *LINGO2*, *MACROD2* and *NKAIN2*). A total of 60 genes associated with 131 CNVs were identified in multiple individuals (as shown in table 3). 52 genes were affected by a CNV in two individuals; five genes were affected by a CNV in three individuals (*IGSF11*, *GK5*, *XRN1*, *NAMPT* and *LCP1*); and three genes were affected by a CNV in four individuals (*CTNNA3*, *NRG3* and *LOC642597*).

While this study has not investigated further the contribution of any one of these CNVs to disease development, previous studies have reported the involvement of several of the genes influenced by one or more CNVs in CRC: L-plastin (subunit *LCP1*) has been shown to be unregulated in various solid human tumours and is also known to contribute to CRC progression via its involvement in cell proliferation and invasion and consequently metastasis[33-35]; alpha-catenin (subunit *CTNNA3*) has been reported to show reduced expression in CRC cell lines which has been suggested to facilitate metastasis[36], while another study has reported increased expression of alpha-catenin during adenoma formation via the negative regulation of beta-catenin signalling[37]; the tumour suppressor gene *APC* has been unequivocally associated with the CRC and Familial adenomatous polypsis (FAP)[38-40]; and furthermore, expression of *IGSF11* has been reported to be elevated in CRC cells lines and may represent a target for cancer immunotherapy[41]. Future studies are required to validate and investigate the role of the CNVs identified in our study for their potential contribution in the development of HNPCC.

Of the 386 CNVs identified unique to the patients, of these regions 56.5% of them have been previously reported in the Database of Genomic Variants (DGV). 59 CNVs

contained genomic regions which were identified in multiple patients (table 4). A total of 15 genomic regions were identified in two patients; five common genomic regions were identified in three patients, located on chromosomes 3, 5, 9, 11 and 12; and one genomic region was identified in four patients on chromosome 16. Two other CNVs were also shown to be common to five patients on chromosomes 3 and 5. Additional studies are required to investigate the sequence content of these regions to identify if novel contributors to disease development may reside in these regions.

Table 3 Genes associated with unique CNVs (compared to controls) identified across multiple among patients. Number of CNV events in which gene (s) have been identified and if they were a CN gain or loss.

Type	2 CNV events		3 CNV events	4 CNV events
Gains	*ADARB2*	*ITGA1*	*GK5*	*LOC642597*
	APC	*KIAA1680*	*IGSF11*	
	ARHGAP19	*LATS2*	*LCP1*	
	B2M	*MLL*	*XRN1*	
	BBOX1	*MSI2*	*NAMPT*	
	C10orf139	*NRSN2*		
	C14orf23	*NXPH1*		
	C20orf96	*ODZ4*		
	C3orf33	*PELO*		
	CNTN5	*PHC3*		
	CNTNAP2	*PRKCI*		
	CSNK2A1	*PSG10*		
	DEFB125	*PSG8*		
	DEFB126	*RBCK1*		
	DEFB127	*RNF125*		
	DEFB128	*RNF138*		
	DEFB129	*SOX12*		
	DEFB132	*TBC1D20*		
	EPHA7	*TFG*		
	FAM134B	*TRIB3*		
	FOXG1	*TRIM69*		
	GPR128	*WDR37*		
	GPR160	*ZCCHC3*		
	GYPE	*ZMYND11*		
Both				*NRG3*
Losses	*CNTN4*			*CTNNA3*
	DCDC1			
	PPP2R3C			
	KIAA0391			

Table 4 Genomic regions associated with unique CNVs (compared to controls) identified across multiple among patients. Note CNV frequency and CNV type (*loss **gain); CNV location (chromosome, start bp and end bp) and size; as well as the confidence score associated with CNV call and the number of probes used to call the CNV are also noted.

Chr	Start (bp)	End (bp)	Size (Kb)	Conf	Probes
2 CNV gains					
3	189,058,439	189,098,718	40.28	0.93	31
3	189,069,317	189,088,009	18.69	0.94	26
4	44,664,798	44,699,744	34.95	0.92	41
4	44,664,798	44,699,744	34.95	0.90	41
8	120,414,388	120,438,172	23.78	0.91	27
8	120,419,721	120,451,773	32.05	0.91	30
11	29,547,229	29,593,722	46.49	0.90	39
11	29,547,756	29,593,722	45.97	0.91	37
16	25,330,672	25,438,375	107.70	0.92	46
16	25,330,672	25,438,375	107.70	0.92	46
3 CNV gains					
3	19,014,033	19,041,376	27.34	0.90	31
3	19,016,875	19,041,376	24.50	0.91	28
3	19,016,875	19,041,376	24.50	0.91	28
5	59,744,695	59,807,906	63.21	0.92	52
5	59,744,695	59,811,770	67.08	0.93	54
5	59,749,693	59,807,906	58.21	0.92	51
9	103,982,826	104,016,588	33.76	0.91	27
9	103,982,826	104,017,715	34.89	0.90	28
9	103,991,205	104,017,715	26.51	0.91	26
11	15,765,333	15,791,331	26.00	0.90	30
11	15,770,233	15,796,302	26.07	0.92	30
11	15,776,946	15,795,665	18.72	0.91	24
12	16,469,855	16,503,960	34.11	0.91	33
12	16,469,855	16,503,960	34.11	0.91	33
12	16,476,470	16,506,851	30.38	0.92	33
4 CNV gains					
16	63,364,955	63,389,659	24.70	0.91	33
16	63,369,029	63,389,029	20.00	0.92	30

16	63,369,960	63,388,189	18.23	0.90	28
16	63,371,038	63,397,352	26.31	0.92	38
5 CNV gains					
5	116,651,923	116,698,621	46.70	0.91	36
5	116,655,439	116,692,153	36.71	0.92	27
5	116,656,039	116,695,730	39.69	0.91	29
5	116,660,694	116,697,347	36.65	0.93	28
5	116,660,694	116,693,035	32.34	0.91	24
2 CNV losses					
1	82,801,000	82,821,932	20.93	0.93	31
1	82,801,000	82,821,932	20.93	0.94	31
2	22,087,558	22,261,901	174.34	0.91	110
2	22,087,558	22,261,901	174.34	0.93	110
2	215,167,158	215,204,595	37.44	0.93	49
2	215,167,158	215,204,595	37.44	0.91	49
3	6,562,398	6,603,706	41.31	0.94	42
3	6,562,398	6,603,706	41.31	0.93	42
3	166,523,809	166,565,186	41.38	0.95	39
3	166,525,250	166,565,186	39.94	0.93	38
5	61,460,851	61,504,678	43.83	0.95	31
5	61,460,851	61,504,678	43.83	0.93	31
7	92,319,307	92,343,906	24.60	0.94	26
7	92,319,307	92,343,906	24.60	0.94	26
9	104,331,902	104,396,632	64.73	0.96	35
9	104,331,902	104,396,632	64.73	0.96	35
5 CNV losses					
3	177,370,126	177,396,832	26.71	0.93	26
3	177,370,126	177,396,832	26.71	0.94	26
3	177,370,126	177,396,832	26.71	0.96	26
3	177,370,126	177,399,625	29.50	0.93	27
3	177,370,126	177,396,832	26.71	0.93	26
2 CNV gain and loss					
7*	110,748,452	111,047,157	298.71	0.93	291
7**	111,007,466	111,052,498	45.03	0.94	25
3**	21,228,980	21,313,310	84.33	0.90	88
3*	21,273,619	21,339,035	65.42	0.92	62

Pathway analysis

WebGestalt[26] pathway analysis software was then used to compare a list of 317 genes associated with CNVs uniquely identified across all patients (compared to controls) to all genes in the human genome (supplementary table 6). Enrichment analysis of KEGG pathways and miR targets was conducted.

KEGG analysis revealed a total of 18 significant pathways in which genes uniquely identified in the patients were enriched (table 5). The most significant pathways identified included those of the carbohydrate digestion and absorption (*p=0.00*12); starch and sucrose metabolism (*p=0.0017*); and metabolic pathways (*p=0.0023*) affecting a total of 11 patients. Previous studies have suggested that changes occurring in metabolic pathways are commonly observed during carcinogenesis and tumour growth[42,43]. In the context of this study, these results suggest the potential existence of a germline predisposition in the affected patients which lead to metabolic conditions that promote disease development. The tight junction pathway (*p=0.0058*) and neurotrophin signalling pathway (*p=0.0058*) were also identified to be enriched and have been shown to play a role in gut permeability and motility[44-46]. These pathways have been well documented for their contribution to CRC[47-50]. It is interesting to also note among the enriched KEGG pathways the prostate cancer pathway (*p=0.0251*) and endometrial cancer pathway (*p=0.0251*) also featured and represent two cancers commonly arising in the general population and in the setting of HNPCC/LS[51-53]. Overall, our KEGG results suggest the existence of genetic risk factors which may act to promote the development of cancer.

Enrichment analysis for targets of miRs identified 65 significant regions within the 3'UTR of the CNV impacted genes unique in the patients. We identified 114 miRs (supplementary table 7) that target these genes regions with over 35% of these having previously been reported to have associations with CRC[54-76]. Of the top 10 most significant regions, 40% of the miRs we identified have been associated with CRC (*miR-141*, mi*R-15A*, mi*R-15B*, mi*R-18A*, mi*R-200A*, mi*R-200B*, mi*R-203*, mi*R-32*, mi*R-429* and mi*R-92*). Overall, our miR enrichment analysis supports reported findings on miR involvement in CRC.

In summary the results obtained from the pathway analysis suggest that many of the genes associated with CNVs uniquely identified in patients are associated with carcinogenesis, tumour growth and disease susceptibility and may be factors in the development of CRC.

Table 5 Enriched KEGG pathways from genes identified from CNVs unique to patients.

KEGG Pathway	Genes in Pathway	Observed	Expected	*p*
Carbohydrate digestion and absorption	44	5	0.32	*0.0012*
Starch and sucrose metabolism	54	5	0.4	*0.0017*
Metabolic pathways	1130	21	8.31	*0.0023*
Salivary secretion	89	5	0.65	*0.0058*
Tight junction	132	6	0.97	*0.0058*
Neurotrophin signaling pathway	127	6	0.93	*0.0058*
Propanoate metabolism	32	3	0.24	*0.0170*
Valine, leucine and isoleucine degradation	44	3	0.32	*0.0251*
Prostate cancer	89	4	0.65	*0.0251*
Ribosome biogenesis in eukaryotes	80	4	0.59	*0.0251*
ErbB signaling pathway	87	4	0.64	*0.0251*
mRNA surveillance pathway	83	4	0.61	*0.0251*
Terpenoid backbone biosynthesis	15	2	0.11	*0.0285*
Malaria	51	3	0.37	*0.0293*
Endometrial cancer	52	3	0.38	*0.0293*
Glycerolipid metabolism	50	3	0.37	*0.0293*
Olfactory transduction	388	8	2.85	*0.0346*
beta-Alanine metabolism	22	2	0.16	*0.0439*

MicroRNA annotation

Identification of miRs that have associations with CRC and processes leading to carcinogenesis may further suggest the involvement of potential target genes in disease development. TAM (a Tool for Annotations of miRs) software[27] was then used to identify meaningful miR categories among the 114 miRs that target significantly enriched 3'UTR regions identified in the patients from previous pathway analysis (table 6). We identified a total of 261 miR categories: 22 families, 33 clusters, 39 functional categories, 162 HMDD and 5 tissue specificity categories. It was identified that miRs were enriched in the family category miR-17 (*p=0.0011*). A total of 10 functional categories were enriched including those associated with onco-miRs (*p=0.0264*), processes of apoptosis (*p=0.0291*) and cell-cycle (*p=0.0406*). For the HMDD category miRs were enriched in various forms of cancer, with cancer enrichment alone accounting for 80% of the most significant findings.

In the context of our study it was reassuring to note the presence of adenocarcinoma (*p=0.0000155*) and colorectal neoplasm's (*p=0.0253*) among the cancers enriched in miR annotation. Cardiovascular diseases (4%), psychological disorders (10%) and infection (4%) represented the minority of other significant finds (all with *p<0.0047*). According to tissue specificity, the placenta represented the most significant miR enriched tissue (*p=0.00001284*). Previous studies have suggested that processes of angiogenesis and vascularisation occurring during placental development in pregnancy are also present during tumour development and this has been observed in CRC[77-79].

Overall the results obtained from the TAM analysis suggest that the 114 miRs associated with the 3'UTR regions significantly identified in the patients may stimulate processes leading to carcinogenesis which is consistent with what we expect to find in these cancer patients[80].

Table 6 Summary of significant findings from miR annotation analysis in TAM software.

Category	# sub-categories	# of Sig. Categories	Significant Findings
Family	22	1	miR-17
Cluster	33	0	-
Function	39	8	Onco-miRs, Apoptosis, Cell cycle related
HMDD	162	20	Cancer (80%), Cardiovascular (4%), Infection (4%) and Psychological disorders (10%)
Tissue	5	1	Placenta

CNV burden

A study by Girirajan and Eichler[81] has suggested that the severity of disease may be explained by the overall burden CNVs place on an individual's genome where increased sensitivity to developing disease is correlated with increased CNV burden and furthermore that variation in CNV burden will result in phenotype variation in patients. In a recent study looking at both HNPCC MMR mutation negative and MMR mutation positive patients, we observed an increased average size of CNVs in patients tested and suggested that this was related to an increased genomic burden[82]. An increased CNV burden was not observed among the patients utilized in the current study, though we did detect a decreased mean size of CNVs in patients compared to controls. Importantly, the current study compared 125 patients and 40 controls whereas the recent HNPCC study compared 96 patients and 384 controls[82]. We suggest that discrepancies in these findings are likely to be related to the inequity of sample populations between studies, the limited number of controls used in the current study, the type of array used (noting differences in both the array coverage and density), as well as the algorithm used by analysis software may all contribute to variation in the observed results[83-85].

CNV bias

CNV analysis has suffered from a lack of standardization in analytical techniques used for data mining. The Hidden Markov Model (HMM) and Circular Binary Segmentation (CBS) represent the algorithms utilized to develop CNV calling programs that have been reported to be the most efficient[83-85]. Furthermore, using algorithms developed for a specific data type has been shown to perform better in CNV calling compared to platform-independent software algorithms[86]. The robustness of software algorithms, batch effects, and population stratification will therefore influence the accuracy of calls made to segmented data and hence the reliability of CNV calls and CNV boundary descriptors derived from arrays[83-85]. The Cyto2.7M array was chosen for use in the current study as at the time it provided the greatest density and most even genomic coverage of any CNV arrays. All data was analysed through ChAS which uses a HMM-based algorithm and was specifically developed to use with the Cyto2.7M array data. While a recent report has suggested using a minimum of three different algorithms when conducting association analysis, this was not possible due to limitations in the data generated by the array used in this study[87].

Conclusions

We were unable to identify any DNA mismatch repair genes targeted by CNVs that may contribute to a significant proportion of HNPCC patients recruited into this study. We did identify several genomic regions that were altered in multiple unrelated HNPCC patients that could potentially be associated with disease risk. The genomic regions encompassed by these CNVs warrant further study to define precisely their role in disease development. We could not rule out the existence of CNVs, smaller than the limits of detection provided by this array, from involvement in the aetiology of HNPCC.

Pathway analysis was thus utilized to identify possible common pathways associated with the heterogeneous outcomes of the analysis. We identified a total of 317 genes impacted by CNVs uniquely identified in patients (compared to controls). Results from KEGG pathway analysis identified the enrichment of pathways involved in metabolism, and these are known to be required for cancer development. It is likely that these loci may contribute to CRC disease risk in the affected individuals. miR enrichment analysis has further highlighted a series of miRs which are suggested to contribute to carcinogenesis. It was found that over 40% of these miRs had been previously reported to play a role in CRC development. As such we have shown that CNV altered genes are over represented in pathways leading to carcinogenesis, tumour growth and disease susceptibility, including CRC. The genes driving pathway enrichments require further investigation to elucidate their precise role in disease development.

The annotation of 114 miRs (reported in the pathway analysis) identified significant functional miR categories associated with cancer, including specifically adenocarcinomas and colorectal neoplasms. Placental tissue was identified to be among the tissues most significantly enriched with the miR looked at in this investigation. We speculate that processes of angiogenesis and vascularisation necessary for placental development are also present during tumour formation including those observed in CRC. As such the genes associated with CNVs we have identified are targeted by miRs which are implicated in various processes leading to malignancy. We conclude that while we have not shown direct consequences of miRs interacting with our CNV altered genes, the separate effects of aberrant miR expression and CNVs impacting on genes that such miRs target may have similar consequences.

Overall the results of this study provide some evidence of CNV involvement in the aetiology of HNPCC and furthermore reinforce that CNV probe arrays compared to SNP arrays appear to be of limited utility for CNV detection.

Conflict of Interest

The authors declare no conflict of interest.

Acknowledgements and declaration of interest

This work has been supported by the following funding bodies and institutions: Australian Rotary Health/Rotary District 9650, the Commonwealth Scientific and Industrial Research Organization (CSIRO), the University of Newcastle and the Hunter Medical Research Institute. Samples were provided by the Hunter Area Pathology Service and the Hunter Community Study.

References

1. Vasen, H.F., Mecklin, J.P., Khan, P.M. & Lynch, H.T. The International Collaborative Group on Hereditary Non-Polyposis Colorectal Cancer (ICG-HNPCC). *Dis Colon Rectum* **34**, 424-5 (1991).

2. Kemp, Z., Thirlwell, C., Sieber, O., Silver, A. & Tomlinson, I. An update on the genetics of colorectal cancer. *Hum Mol Genet* **13 Spec No 2**, R177-85 (2004).

3. Peltomaki, P. Deficient DNA mismatch repair: a common etiologic factor for colon cancer. *Hum Mol Genet* **10**, 735-40 (2001).

4. Thompson, E. *et al.* Hereditary non-polyposis colorectal cancer and the role of hPMS2 and hEXO1 mutations. *Clin Genet* **65**, 215-25 (2004).

5. Kuiper, R.P. *et al.* Recurrence and variability of germline EPCAM deletions in Lynch syndrome. *Hum Mutat* **32**, 407-14 (2011).

6. Ligtenberg, M.J. *et al.* 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 (2009).

7. McPhillips, M., Meldrum, C.J., Creegan, R., Edkins, E. & Scott, R.J. Deletion Mutations in an Australian Series of HNPCC Patients. *Hered Cancer Clin Pract* **3**, 43-7 (2005).

8. Bonis, P.A. *et al.* Hereditary nonpolyposis colorectal cancer: diagnostic strategies and their implications. *Evid Rep Technol Assess (Full Rep)*, 1-180 (2007).

9. Obermair, A. *et al.* Risk of Endometrial Cancer for women diagnosed with HNPCC-related colorectal cancer. *International journal of cancer* **127**, 7 (2010).

10. Ionita-Laza, I., Rogers, A.J., Lange, C., Raby, B.A. & Lee, C. Genetic association analysis of copy-number variation (CNV) in human disease pathogenesis. *Genomics* **93**, 22-6 (2009).

11. Almal, S.H. & Padh, H. Implications of gene copy-number variation in health and diseases. *J Hum Genet* **57**, 6-13 (2012).

12. Bronstad, I., Wolff, A.S., Lovas, K., Knappskog, P.M. & Husebye, E.S. Genome-wide copy number variation (CNV) in patients with autoimmune Addison's disease. *BMC Med Genet* **12**, 111 (2011).

13. Grozeva, D. *et al.* Rare copy number variants: a point of rarity in genetic risk for bipolar disorder and schizophrenia. *Arch Gen Psychiatry* **67**, 318-27 (2010).

14. Hai, R. *et al.* Genome-wide association study of copy number variation identified gremlin1 as a candidate gene for lean body mass. *J Hum Genet* **57**, 33-7 (2012).

15. Jiang, Q., Ho, Y.Y., Hao, L., Nichols Berrios, C. & Chakravarti, A. Copy number variants in candidate genes are genetic modifiers of Hirschsprung disease. *PLoS One* **6**, e21219 (2011).

16. Wellcome Trust Case Control, C. *et al.* Genome-wide association study of CNVs in 16,000 cases of eight common diseases and 3,000 shared controls. *Nature* **464**, 713-20 (2010).

17. Delnatte, C. *et al.* Contiguous gene deletion within chromosome arm 10q is associated with juvenile polyposis of infancy, reflecting cooperation between the BMPR1A and PTEN tumor-suppressor genes. *Am J Hum Genet* **78**, 1066-74 (2006).

18. van Hattem, W.A. *et al.* Large genomic deletions of SMAD4, BMPR1A and PTEN in juvenile polyposis. *Gut* **57**, 623-7 (2008).

19. Fokkema, I.F. *et al.* LOVD v.2.0: the next generation in gene variant databases. *Hum Mutat* **32**, 557-63 (2011).

20. Nagasaka, T. *et al.* Somatic hypermethylation of MSH2 is a frequent event in Lynch Syndrome colorectal cancers. *Cancer Res* **70**, 3098-108 (2010).

21. McEvoy, M. *et al.* Cohort profile: The Hunter Community Study. *Int J Epidemiol* **39**, 1452-63 (2010).

22. Miller, S.A., Dykes, D.D. & Polesky, H.F. A simple salting out procedure for extracting DNA from human nucleated cells. *Nucleic Acids Res* **16**, 1215 (1988).

23. Vasen, H.F., Watson, P., Mecklin, J.P. & Lynch, H.T. New clinical criteria for hereditary nonpolyposis colorectal cancer (HNPCC, Lynch syndrome) proposed by the International Collaborative group on HNPCC. *Gastroenterology* **116**, 1453-6 (1999).

24. Rodriguez-Bigas, M.A. *et al.* A National Cancer Institute Workshop on Hereditary Nonpolyposis Colorectal Cancer Syndrome: meeting highlights and Bethesda guidelines. *J Natl Cancer Inst* **89**, 1758-62 (1997).

25. QuickCalcs - T test. (GraphPad Software Inc., GraphPad Software Inc., 2013).

26. Zhang, B., Kirov, S. & Snoddy, J. WebGestalt: an integrated system for exploring gene sets in various biological contexts. *Nucleic Acids Res* **33**, W741-8 (2005).

27. Lu, M., Shi, B., Wang, J., Cao, Q. & Cui, Q. TAM: a method for enrichment and depletion analysis of a microRNA category in a list of microRNAs. *BMC Bioinformatics* **11**, 419 (2010).

28. Feuk, L., Carson, A.R. & Scherer, S.W. Structural variation in the human genome. *Nat Rev Genet* **7**, 85-97 (2006).

29. Sebat, J. *et al.* Large-scale copy number polymorphism in the human genome. *Science* **305**, 525-8 (2004).

30. Sharp, A.J. *et al.* Segmental duplications and copy-number variation in the human genome. *Am J Hum Genet* **77**, 78-88 (2005).

31. Krepischi, A.C. *et al.* Germline DNA copy number variation in familial and early-onset breast cancer. *Breast Cancer Res* **14**, R24 (2012).

32. Mitchell, R.J., Farrington, S.M., Dunlop, M.G. & Campbell, H. Mismatch repair genes hMLH1 and hMSH2 and colorectal cancer: a HuGE review. *Am J Epidemiol* **156**, 885-902 (2002).

33. Foran, E., McWilliam, P., Kelleher, D., Croke, D.T. & Long, A. The leukocyte protein L-plastin induces proliferation, invasion and loss of E-cadherin expression in colon cancer cells. *Int J Cancer* **118**, 2098-104 (2006).

34. Otsuka, M. *et al.* Differential expression of the L-plastin gene in human colorectal cancer progression and metastasis. *Biochem Biophys Res Commun* **289**, 876-81 (2001).

35. Park, T., Chen, Z.P. & Leavitt, J. Activation of the leukocyte plastin gene occurs in most human cancer cells. *Cancer Res* **54**, 1775-81 (1994).

36. Sygut, A. *et al.* Genetic Variations of the CTNNA1 And The CTNNB1 Genes in Sporadic Colorectal Cancer in Polish Population. *Pol Przegl Chir* **84**, 560-4 (2012).

37. Giannini, A.L., Vivanco, M. & Kypta, R.M. alpha-catenin inhibits beta-catenin signaling by preventing formation of a beta-catenin*T-cell factor*DNA complex. *J Biol Chem* **275**, 21883-8 (2000).

38. Patel, S.G. & Ahnen, D.J. Familial colon cancer syndromes: an update of a rapidly evolving field. *Curr Gastroenterol Rep* **14**, 428-38 (2012).

39. Albuquerque, C. *et al.* Colorectal cancers show distinct mutation spectra in members of the canonical WNT signaling pathway according to their anatomical location and type of genetic instability. *Genes Chromosomes Cancer* **49**, 746-59 (2010).

40. Christie, M. *et al.* Different APC genotypes in proximal and distal sporadic colorectal cancers suggest distinct WNT/beta-catenin signalling thresholds for tumourigenesis. *Oncogene* (2012).

41. Watanabe, T. *et al.* Identification of immunoglobulin superfamily 11 (IGSF11) as a novel target for cancer immunotherapy of gastrointestinal and hepatocellular carcinomas. *Cancer Sci* **96**, 498-506 (2005).

42. DeBerardinis, R.J. & Thompson, C.B. Cellular metabolism and disease: what do metabolic outliers teach us? *Cell* **148**, 1132-44 (2012).

43. Munoz-Pinedo, C., El Mjiyad, N. & Ricci, J.E. Cancer metabolism: current perspectives and future directions. *Cell Death Dis* **3**, e248 (2012).

44. Joo, Y.E. Increased Expression of Brain-derived Neurotrophic Factor in Irritable Bowel Syndrome and Its Correlation With Abdominal Pain (Gut 2012;61:685-694). *J Neurogastroenterol Motil* **19**, 109-11 (2013).

45. Ulluwishewa, D. *et al.* Regulation of tight junction permeability by intestinal bacteria and dietary components. *J Nutr* **141**, 769-76 (2011).

46. Visser, J., Rozing, J., Sapone, A., Lammers, K. & Fasano, A. Tight junctions, intestinal permeability, and autoimmunity: celiac disease and type 1 diabetes paradigms. *Ann N Y Acad Sci* **1165**, 195-205 (2009).

47. Akil, H., Perraud, A., Melin, C., Jauberteau, M.O. & Mathonnet, M. Fine-tuning roles of endogenous brain-derived neurotrophic factor, TrkB and sortilin in colorectal cancer cell survival. *PLoS One* **6**, e25097 (2011).

48. Enam, S., Gan, D.D., White, M.K., Del Valle, L. & Khalili, K. Regulation of human neurotropic JCV in colon cancer cells. *Anticancer Res* **26**, 833-41 (2006).

49. Wang, X., Tully, O., Ngo, B., Zitin, M. & Mullin, J.M. Epithelial tight junctional changes in colorectal cancer tissues. *ScientificWorldJournal* **11**, 826-41 (2011).

50. Soler, A.P. *et al.* Increased tight junctional permeability is associated with the development of colon cancer. *Carcinogenesis* **20**, 1425-31 (1999).

51. Grindedal, E.M. *et al.* Germ-line mutations in mismatch repair genes associated with prostate cancer. *Cancer Epidemiol Biomarkers Prev* **18**, 2460-7 (2009).

52. Desai, M.D., Saroya, B.S. & Lockhart, A.C. Investigational therapies targeting the ErbB (EGFR, HER2, HER3, HER4) family in GI cancers. *Expert Opin Investig Drugs* **22**, 341-56 (2013).

53. Khelwatty, S.A., Essapen, S., Seddon, A.M. & Modjtahedi, H. Prognostic significance and targeting of HER family in colorectal cancer. *Front Biosci* **18**, 394-421 (2013).

54. Baraniskin, A. *et al.* MiR-30a-5p suppresses tumor growth in colon carcinoma by targeting DTL. *Carcinogenesis* **33**, 732-9 (2012).

55. Bauer, K.M. & Hummon, A.B. Effects of the miR-143/-145 microRNA cluster on the colon cancer proteome and transcriptome. *J Proteome Res* **11**, 4744-54 (2012).

56. Cekaite, L. *et al.* MiR-9, -31, and -182 deregulation promote proliferation and tumor cell survival in colon cancer. *Neoplasia* **14**, 868-79 (2012).

57. Cheng, H. *et al.* Circulating plasma MiR-141 is a novel biomarker for metastatic colon cancer and predicts poor prognosis. *PLoS One* **6**, e17745 (2011).

58. Dai, L. *et al.* Vector-based miR-15a/16-1 plasmid inhibits colon cancer growth in vivo. *Cell Biol Int* **36**, 765-70 (2012).

59. Franke, A. *et al.* Genome-wide meta-analysis increases to 71 the number of confirmed Crohn's disease susceptibility loci. *Nat Genet* **42**, 1118-25 (2010).

60. He, X. *et al.* MicroRNA-218 inhibits cell cycle progression and promotes apoptosis in colon cancer by downregulating oncogene BMI-1. *Mol Med* (2012).

61. Kenny, E.E. *et al.* A genome-wide scan of Ashkenazi Jewish Crohn's disease suggests novel susceptibility loci. *PLoS Genet* **8**, e1002559 (2012).

62. Migliore, C. *et al.* MiR-1 downregulation cooperates with MACC1 in promoting MET overexpression in human colon cancer. *Clin Cancer Res* **18**, 737-47 (2012).

63. Nie, J. *et al.* microRNA-365, down-regulated in colon cancer, inhibits cell cycle progression and promotes apoptosis of colon cancer cells by probably targeting Cyclin D1 and Bcl-2. *Carcinogenesis* **33**, 220-5 (2012).

64. Okamoto, K. *et al.* miR-493 induction during carcinogenesis blocks metastatic settlement of colon cancer cells in liver. *EMBO J* **31**, 1752-63 (2012).

65. Okayama, H., Schetter, A.J. & Harris, C.C. MicroRNAs and inflammation in the pathogenesis and progression of colon cancer. *Dig Dis* **30 Suppl 2**, 9-15 (2012).

66. Qased, A.B. *et al.* MicroRNA-18a upregulates autophagy and ataxia telangiectasia mutated gene expression in HCT116 colon cancer cells. *Mol Med Report* **7**, 559-64 (2013).

67. Roy, S., Levi, E., Majumdar, A.P. & Sarkar, F.H. Expression of miR-34 is lost in colon cancer which can be re-expressed by a novel agent CDF. *J Hematol Oncol* **5**, 58 (2012).

68. Slaby, O., Svoboda, M., Michalek, J. & Vyzula, R. MicroRNAs in colorectal cancer: translation of molecular biology into clinical application. *Mol Cancer* **8**, 102 (2009).

69. Strillacci, A. *et al.* Loss of miR-101 expression promotes Wnt/beta-catenin signalling pathway activation and malignancy in colon cancer cells. *J Pathol* **229**, 379-89 (2013).

70. Sun, J.Y. *et al.* MicroRNA-320a suppresses human colon cancer cell proliferation by directly targeting beta-catenin. *Biochem Biophys Res Commun* **420**, 787-92 (2012).

71. Wang, Z. *et al.* MiR-145 regulates PAK4 via the MAPK pathway and exhibits an antitumor effect in human colon cells. *Biochem Biophys Res Commun* **427**, 444-9 (2012).

72. Weissmann-Brenner, A. *et al.* Tumor microRNA-29a expression and the risk of recurrence in stage II colon cancer. *Int J Oncol* **40**, 2097-103 (2012).

73. Wu, J. *et al.* Up-regulation of microRNA-1290 impairs cytokinesis and affects the reprogramming of colon cancer cells. *Cancer Lett* **329**, 155-63 (2013).

74. Wu, J. *et al.* MicroRNA-34a inhibits migration and invasion of colon cancer cells via targeting to Fra-1. *Carcinogenesis* **33**, 519-28 (2012).

75. Zhang, J. *et al.* miR-21, miR-17 and miR-19a induced by phosphatase of regenerating liver-3 promote the proliferation and metastasis of colon cancer. *Br J Cancer* **107**, 352-9 (2012).

76. Zhu, R. *et al.* Ascl2 knockdown results in tumor growth arrest by miRNA-302b-related inhibition of colon cancer progenitor cells. *PLoS One* **7**, e32170 (2012).

77. Harada, O. *et al.* The role of trophinin, an adhesion molecule unique to human trophoblasts, in progression of colorectal cancer. *Int J Cancer* **121**, 1072-8 (2007).

78. Hatakeyama, K. *et al.* Placenta-specific novel splice variants of Rho GDP dissociation inhibitor beta are highly expressed in cancerous cells. *BMC Res Notes* **5**, 666 (2012).

79. Wei, S.C. *et al.* Placenta growth factor expression is correlated with survival of patients with colorectal cancer. *Gut* **54**, 666-72 (2005).

80. Hrasovec, S. & Glavac, D. MicroRNAs as Novel Biomarkers in Colorectal Cancer. *Front Genet* **3**, 180 (2012).

81. Girirajan, S. & Eichler, E.E. Phenotypic variability and genetic susceptibility to genomic disorders. *Hum Mol Genet* **19**, R176-87 (2010).

82. Talseth-Palmer, B.A. *et al.* Continuing difficulties in interpreting CNV data: lessons from a genome-wide CNV association study of Australian HNPCC/lynch syndrome patients. *BMC Med Genomics* **6**, 10 (2013).

83. Dellinger, A.E. *et al.* Comparative analyses of seven algorithms for copy number variant identification from single nucleotide polymorphism arrays. *Nucleic Acids Res* **38**, e105 (2010).

84. Tsuang, D.W. *et al.* The effect of algorithms on copy number variant detection. *PLoS One* **5**, e14456 (2010).

85. Zhang, D. *et al.* Accuracy of CNV Detection from GWAS Data. *PLoS One* **6**, e14511 (2011).

86. Pinto, D. *et al.* Comprehensive assessment of array-based platforms and calling algorithms for detection of copy number variants. *Nat Biotechnol* **29**, 512-20 (2011).

87. Kim, S.Y., Kim, J.H. & Chung, Y.J. Effect of Combining Multiple CNV Defining Algorithms on the Reliability of CNV Calls from SNP Genotyping Data. *Genomics Inform* **10**, 194-9 (2012).

CHAPTER 4: DEEP INTRONIC VARIANTS RESULTING IN ABERRANT MRNA SPECIES IN CONTRIBUTION TO HNPCC

Introduction

Hereditary non-polyposis colorectal cancer (HNPCC) represents the most common predisposition to colorectal cancer and is unequivocally associated with defects in one of four DNA mismatch repair (MMR) genes *MLH1*, *MSH2*, *MSH6* or *PMS2*[63-69]. Approximately 50% of mutations detected are identified in *MLH1*, 40% in *MSH2*, 5-10% in *MSH6* and a few families have *PMS2* mutations[143,147,182]. Collectively however, mutations in these genes only account for 50% of clinically tested patients suggesting that other genes or mechanisms of gene silencing could be responsible for LS.

Mutation screening, for HNPCC in the clinical setting, typically involved Sanger sequencing for sequence variants and Multiplex ligation-dependant probe amplification (MLPA) for the detection of duplications and deletions. Both techniques are highly targeted and do not usually encompass deep intronic regions of the target genes among the genomic sequence being tested. Despite this, aberrations in non-coding regions of genes are known to give rise to disease[350-353].

This part of the thesis aims to further define the role and frequency of deep intronic mutations in *MLH1*, *MSH2* and *MSH6* in contribution to disease in a series of MMR mutation negative HNPCC patients.

Publication

Amy L. Masson, Bente A. Talseth-Palmer, Tiffany-Jane Evans, Trish Collinson, Michelle Wong-Brown, Melissa A. Tooney, Garry N. Hannan and Rodney J. Scott (2015) Intronic variants resulting in aberrant mRNA species are rare in Hereditary Non-Polyposis Colorectal Cancer, *Hereditary Cancer in Clinical Practice*.

Co-author statement

I attest that Research Higher Degree candidate Amy Louise Masson contributed to the above manuscript including involvement in the conception and design of the study; conducting the laboratory work, data analysis and interpretation; and preparation of the manuscript.

Co-author	Signature	Date
Bente A. Talseth-Palmer		
Tiffany-Jane Evans		
Trish Collinson		
Michelle Wong-Brown		
Melissa A. Tooney		
Garry N. Hannan		
Rodney J. Scott		

Amy Louise Masson Date: 01/09/2015

Professor Robert Callister Date: 01/09/2015

Assistant Dean Research Training

Intronic variants resulting in aberrant mRNA species are rare in Hereditary Non-Polyposis Colorectal Cancer

Amy L. Masson[1,2], Bente A. Talseth-Palmer[1,2], Tiffany-Jane Evans[1,2], Trish Collinson[1,2], Michelle Wong-Brown[1,2], Melissa Anne Tooney[3], Garry N. Hannan[4] and Rodney J. Scott[1,2,3,*]

[1] Information Based Medicine Program, Hunter Medical Research Institute, University of Newcastle, Newcastle, New South Wales, 2305 Australia; E-Mails: Amy.L.Masson@uon.edu.au (A.L.M.); Bente.Talseth-Palmer@newcastle.edu.au (B.A.T-P.); Tiffany-Jane.Evans@newcastle.edu.au (T-J.E.); trish.collinson@newcastle.edu.au (T.C.); Michelle.Wong-Brown@newcastle.edu.au (M.W-B)

[2] School of Biomedical Sciences and Pharmacy, Faculty of Health, University of Newcastle, New South Wales, 2308 Australia; Emails: Melissa.Tooney@hnehealth.nsw.gov.au (M.A.T)

[3] Division of Molecular Medicine, Hunter Area Pathology Service, John Hunter Hospital, Newcastle, New South Wales, 2305 Australia; [4] CSIRO Food and Nutrition Flagship, North Ryde, New South Wales, 2113 Australia; Emails: Garry.Hannan@csiro.au (G.N.H.)

*Author to whom correspondence should be addressed; E-mail: Rodney.Scott@newcastle.edu.au (R.J.S.); Tel.: +61 (2) 4921 4974; Fax: +61 (2) 4921 4253

Abstract

Background

Hereditary non-polyposis colorectal cancer (HNPCC) accounts for the largest proportion of patients with a genetic predisposition to develop colorectal cancer (CRC) at unusually young ages. Germline mutations in one of four mismatch repair (MMR) genes (*MLH1*, *MSH2*, *MSH6* and *PMS2)* form a large subset of HNPCC known as Lynch syndrome (LS). In patients diagnosed with HNPCC that have been shown not to harbour deleterious changes in the exons of any of the four MMR genes there remains the possibility that more cryptic changes occur in intronic regions. Here we have screened forty-six HNPCC MMR mutation negative patients for variants contained within the intronic regions of the MMR genes *MLH1*, *MSH2* and *MSH6* which may contribute to disease via the formation of cryptic splice sites and the formation of pseudo exons.

Methods

Lymphoblastoid cells (B-lymphocytes) were obtained from forty-six HNPCC patients and immortalized using EBV for RNA extraction. RNA was reverse transcribed into cDNA and assayed by size fractionation of thirteen fragments covering the exon-exon boundaries of *MLH1*, *MSH2* and *MSH6*. Products that were found to be smaller or larger than the expected reference sequence were sequenced by Sanger sequencing and any variants identified by mutation surveyor checked against the LOVD database and referenced according to the InSight classification of pathogenicity.

Results

Of the forty-six cell lines (representing the 46 patients), two were identified that harboured non-pathogenic variants which resided in the flanking exonic sequences of the respective gene fragments screened.

Conclusions

This study revealed that pathogenic variants residing in intronic regions of the three MMR genes (*MSH2*, *MSH6* and *MLH1*) are rare in patients diagnosed with HNPCC. Future studies should focus on the presence of alternative splice forms of the respective genes and their potential contribution to disease.

Key Words

mRNA transcript variants; HNPCC/Lynch Syndrome; intronic variants; diagnostic testing.

Introduction

Hereditary non-polyposis colorectal cancer (HNPCC) represents somewhere where between 2% and 5% of all colorectal cancers (CRCs) and by definition, describes families who conform to either the Amsterdam Criteria, the Amsterdam Criteria II or the Bethesda Criteria[1], which were originally developed to identify the genetic basis of the disease[2].

A subset of HNPCC patients that are identified to harbour heritable germline alterations that result in the inactivation of one of four DNA mismatch repair (MMR) genes, *MLH1*, *MSH2*, *MSH6* or *PMS2* are now specifically described as having Lynch Syndrome (LS)[3-5]. Although mutations in *MLH1*, *MSH2*, *MSH6* and *PMS2* account for the majority of LS families[3-9], variants affecting *EPCAM* (transcriptional silencing of *MSH2*) have also been implicated in the disease by virtue of the associated silencing of *MSH2*[10,11].

MLH1 is comprised of 19 exons and has a full-length transcript of 2662 base pairs (bp). *MSH2* and *MSH6* are comprised of 16 and 10 exons, respectively. *MSH2* and *MSH6* full-length transcripts are 3226 bp and 4435 bp, respectively, in size. *PMS2* is comprised of 18 exons and has a full length transcript of 2851 bp. A recent review reported the existence of at least 30 *MLH1*, 22 *MSH2*, 4 *MSH6* and 9 *PMS2* naturally occurring splice variants[12-16] that can complicate the interpretation of aberrant transcripts when investigating this disease.

Diagnostic testing for aberrations in the four MMR genes typically involves surveying the exons (and flanking intronic sequence) of the respective genes and undertaking duplication/deletion analysis. Germline mutations in *MLH1* account for approximately 50% of the mutations detected across the four MMR genes with a further 40% identified in *MSH2* and 5-10% in *MSH6*[3,4]. Due to the presence of highly homologous pseudogenes relatively few *PMS2* mutations have been identified[17]. Overall, up to 50% of clinically tested patients with tumours demonstrating microsatellite instability (MSI), a hallmark of HNPCC, fail to have any germline mutation identified in any one of the four MMR genes tested[18-20].

Evidence is growing in support intronic mutations that lead to aberrant mRNA splicing and non-functional gene transcripts in disease development as crucial elements for normal splicing (e.g. donor splice sites, the branch point, the acceptor splice site, the polypyrimidine tracts and splicing enhancers/silencers) may be altered[21]. To date, several reports have described variants in the intronic regions of genes that give rise to alternative splicing patterns which contribute to CRC[22-25]) including a report on 15

HNPCC MMR mutation negative cases carrying intronic mutations in *MLH1* or *MSH2* that found 10 of the 15 base substitutions identified were likely to result in exon skipping[26] and another report on 125 familial adenomatous polyposis (FAP) *APC* mutation negative cases identified aberrant *APC* mRNA transcripts in 8% of their patients and suggest that deep intronic mutations gave rise to cryptic splice sites and the formation of out-of-frame pseudoexons[23].

To further define the role and frequency of occurrence of deep cryptic intronic mutations and their potential contribution to LS, we screened 46 unrelated MMR mutation negative HNPCC patients for intronic variants in the MMR genes *MLH1*, *MSH2* and *MSH6*.

Methods

Samples

The 46 lymphoblastoid cell lines (LCLs) used in the current study were prepared by immortalizing B-lymphocytes isolated from 46 unrelated HNPCC patients who had given informed consent for their samples to be used for studies into their disease. The B-lymphocytes had been stored in liquid nitrogen, from 1998 - 2010. All contributing patients fulfilled either the Amsterdam Criteria, the Amsterdam II Criteria[2,27] or the Bethesda Guidelines[1]. All patients had been diagnosed with CRC and were the first individual (proband) of their family to seek genetic testing. The samples were referred for routine clinical diagnostic testing involving screening for mutations in: *MLH1*, *MSH2*, *MSH6* and/or *PMS2*. Mutation screening was performed using Sanger Sequencing and Multiplex ligation-dependant probe amplification (MLPA) analysis. No mutations were identified in any of the patients used for the current study and were therefore considered to be MMR mutation negative. Average age of diagnosis across this cohort was 41 years. The study was approved by the Hunter New England Human Research Ethics Committee (HNEHREC) and the University of Newcastle's Human Research Ethics Committee (HREC). See table 1 for patient details and immune-histochemical status (i.e. protein expression present or absent in the respective tumour).

Table 1 Clinical information related to 46 HNPCC patients from which 46 lymphoblast cell lines were cultured. Note the Patient ID, date of birth (DOB), age of colorectal cancer diagnosis (Dx), sex (M/F, M=Male and F=Female), genes tested (Sanger sequencing and Multiplex Ligation-dependant Probe Amplification) by clinical lab, variants (non-pathogenic) identified, immunohistochemistry (IHC) test results, other cancers and age (x) diagnosed in patient and finally, the clinical criteria used to diagnose the patient with HNPCC.

Patient	DOB	Dx	M/F	Genes tested	Variants	IHC	Other cancers	Crit
HNPCC_01	07.03.39	48	M	*MLH1* *MSH2* *MSH6*				AMII
HNPCC_02	12.02.41	41	M	*MLH1* *MSH2*		+ve *MLH1*		AMII
HNPCC_03	31.03.34	39	F	*MLH1* *MSH2*				AMII
HNPCC_04	25.05.59	45	F	*MSH6*		-ve *MSH6*		AMII
HNPCC_05	06.06.57	49	F	*MLH1* *PMS2*	*MLH1,* Ex 8, c.655A>G, I219V	-ve *MLH1* *PMS2*	Melanoma (45)	AMII
HNPCC_06	05.04.49	35	M	*MLH1* *MSH2* *MSH6*	*MSH6*, Ex 1, c.116G>A, G39E	-ve *MSH2* *MSH6* +ve *MLH1*	Skin (48); CRC (50), caecum; CRC (54), descending colon	AMII
HNPCC_07	30.12.44	49	M	*MLH1* *MSH2*				AMII
HNPCC_08	14.11.55	42	F	*MLH1* *MSH2*	*MSH2*, Ex 7, c.1077-10T>C, ?; *MSH2*, Ex 13, c.2006-6T>C, ?	-ve *MSH6*	Cholangia (44)	AMII
HNPCC_09	16.09.28	49	M	*MLH1* *MSH2* *MSH6*				AMII
HNPCC_10	01.12.61	41	M	*MSH2* *MSH6*	*MSH6*, Ex 4, c.1870G>A,	- ve *MSH2*		AMII

					G624S			
HNPCC_11	14.09.60	36	F	*MLH1* *MSH2*			Bladder (35); Uterine (35)	AMII
HNPCC_12	23.04.42	31	M	*MLH1* *MSH2*				AMII
HNPCC_13	18.08.63	39	M	*MLH1* *MSH2* *PMS2*		-ve *MLH1* *PMS2* +ve *MSH2* *MSH6*		AMII
HNPCC_14	02.09.51	47	F	*MLH1* *MSH2*				AMII
HNPCC_15	12.12.30	60	F	*MLH1* *MSH2*		+ve *MLH1* *MSH2*		AMII
HNPCC_16	21.06.51	42	F	*MLH1* *MSH2*			Ovarian (55)	AMII
HNPCC_17	11.06.65	37	F	*MLH1* *PMS2*	*MLH1*, Ex 18, c.2038T >C, C680R; *PMS2*, Ex 11, C.1688G >T, Arg563Leu			AMII
HNPCC_18	31.05.51	47	M	*MLH1*				AMII
HNPCC_19	03.01.35	32	F	*MLH1* *MSH2*				AMII
HNPCC_20	01.08.45	45	F	*MLH1* *MSH2*				AMII
HNPCC_21	06.02.56	41	F	*MLH1* *MSH2*				AMII
HNPCC_22	03.11.33	50	M	*MLH1* *MSH2*				AMII
HNPCC_23	31.12.49	50	F	*MLH1* *MSH2*	*MSH2*, Ex 11, c.?, A564T			AMII
HNPCC_24	02.07.60	30	M	*MLH1* *MSH2* *MSH6*	*MSH6*, Ex 4, c.?, K428E	-ve *MSH6* +ve		AMII

						MLH1 *MSH2*		
HNPCC_25	25.12.53	45	F	*MLH1* *MSH2*				AMII
HNPCC_26	28.08.66	32	F	*MLH1* *MSH2*				AMII
HNPCC_27	08.01.58	41	M	*MLH1* *MSH2*				AMII
HNPCC_28	16.02.49	46	M	*MLH1* *MSH2*		+ve *MLH1* *MSH2* *MSH6*	Lung (51)	AMII
HNPCC_29	04.08.62	39	F	*MLH1* *MSH2*		-ve *MSH2*		AMII
HNPCC_30	16.01.59	46	F	*MLH1*		-ve *MLH1*		AMII
HNPCC_31	21.07.34	46	M	*MLH1* *MSH2*				AMII
HNPCC_32	30.07.62	38	F	*MLH1* *MSH2*	*MLH1*, Ex 11, c.931A>G, Lys311Glu			AMII
HNPCC_33	01.06.41	36	F	*MLH1* *MSH2*				AMII
HNPCC_34	04.08.47	50	F	*MLH1* *MSH2* *MSH6*	*MLH1*, Ex 8, c.655A>G, I219V	-ve *MLH1* *MSH2* *MSH6* *PMS2*		AMII
HNPCC_35	27.12.54	47	F	*MLH1* *MSH2*	*MLH1*, Ex 16, c.1852AA>GC, K618A			AMII
HNPCC_36	15.02.82	21	F	*MLH1* *MSH2*				AMII
HNPCC_37	16.10.61	45	F	*MLH1*		-ve *MLH1*		AMII
HNPCC_38	17.06.53	44	F	*MLH1* *MSH2*				AMII
HNPCC_39	14.11.36	48	F	*MLH1* *MSH2*		+ve *MLH1* *MSH2* *MSH6*	CRC (67)	AMII

HNPCC_40	05.09.80	25	M	*MLH1*	*MLH1*, Ex 8, c.655A>G, I219V	-ve *MLH1*		Beth
HNPCC_41	16.02.57	30	F	*MLH1* *MSH2*				AMII
HNPCC_42	06.08.58	46	M	*MLH1* *MSH2*		+ve *MLH1* *MSH2*		AMII
HNPCC_43	14.12.43	46	M	*MLH1* *MSH2*			Melanoma (42)	AMII
HNPCC_44	28.03.52	40	M	*MLH1* *MSH2*	*MLH1*, Ex 8, c.655A>G, I219V; *MLH1,* Ex 11, c.974G>A, Arg325Gln	-ve *MLH1* *PMS2* +ve *MSH2* *MSH6*		AMII
HNPCC_45	24.06.49	38	F	*MLH1* *MSH2*				AMII
HNPCC_46	09.01.55	28	F	*MLH1* *MSH2*				AMII

Cell Culture

The frozen cells were washed in 2 mL RPMI and 400 μL Foetal Bovine Serum (FBS)) and were collected by centrifugation (140 x g, 10 min), resuspended in R- Media (2 mL; composed of 46.5 mL RPMI, 1 mL pen/strep, 1 mL sodium bicarbonate, 1 mL HEPES buffer and 0.5 mL L-Glut) to which EBV supernatant (1.5 mL), FBS (1 mL) and Cyclosporin A (0.5 mL) were added. The mixture was transferred to a cell culture plate and incubated (37°C, 5% CO_2). The cells were fed with R complete media (composed of 41.5 mL RPMI, 5 mL FBS, 1 mL 6pen/strep, 1 mL sodium bicarbonate, 1 mL HEPES buffer and 0.5 mL L-Glut) as needed every 3-4 days. Once the cells showed a high degree of confluence, they were transferred to flasks and expanded until sufficient for harvesting. Cells were harvested as required by centrifugation (140 x g, 10 min), the supernatant discarded and Phosphate Buffering Solution (PBS; 4.5 mL) added to resuspend cells which were then divided across three microfuge tubes.

RNA extraction, DNAse treatment and RNA purification

To each tube TRIsol Reagent (Invitrogen, Carlsbad California USA; 0.5 mL; incubated at RT 5 min) was added followed by chloroform (0.2 mL), shaken vigorously and incubated RT for 3 min) before centrifuging (12,000 × g, 15 min 4^{o}C). Approximately 400 μL upper phase containing the RNA was transferred to a fresh RNase–DNase free tube before mixing with 500 μL 70% ethanol. The RNA sample was then processed using the PureLink RNA Mini Kit (Ambion, Austin Texas USA) in conjunction with the On-column PureLink DNase (Invitrogen) according to manufacturer's instructions (to remove any possible contamination of genomic DNA in the sample). Sample quantity and quality was evaluated using the Epoch Spectrometer at RNA OD_{260} with all isolated and purified RNA deemed acceptable for use with a 260:280 ratio above 1.9.

Reverse transcription

RNA samples (at a concentration of 200 ng/μL) were reverse transcribed using the High Capacity cDNA Reverse Transcription Kit using RNase inhibitor (Applied Biosystems, Forster City California USA) according to manufactures protocols. Briefly, the kit was thawed on ice and components combined (per reaction: 2 μL 10X RT buffer, 0.8 μL 25X dNTP Mix (100 mM), 2 μL 10X RT Random Primers, 1 L MultiScribe RT, 1 μL RN Inhibitor and 3.2 μL water; placed on ice and gently mixed). To a 96-well plate the reverse transcription master mix (10 μL) was added to each RNA sample (2 μg total, being 10 uL@200ng/μL) sealed, mixed and then placed on ice. The samples

were reverse transcribed using a thermal cycler (25°C 10 min, 27°C 120 min, 85°C 5 min, hold 4°C).

Fragment amplification and analysis

Primers (see table 2, figure 1 and figure 2, fragment location for each gene and primer optimization electrophoresis results) were optimized before use. Note that *MLH1* has two amplicons that span exon 1-4 (Fragments F1 and F2) as two common transcript variants exist for exon 1. Several controls for which RNA was derived from whole blood were also screened to ascertain the sensitivity of the assay with respect to the detection of normal splice variants for the three genes (see figure 2). The presence of the splice variants were confirmed by Sanger sequencing. Thereafter, the cDNA from the patients for fragments of *MLH1*, *MSH2* and *MSH6* genes were amplified according to standard PCR protocols using the HotStar Plus Master Mix kit (Qiagen, Venlo Netherlands). Briefly, a master mix was prepared (10 μL Hotstar Plus Master Mix, 4 μL water, 2 μL each primer) was added to the sample cDNA (2 μL at 200 ng/μL, total 400 ng transcribed RNA) and amplified using a thermal cycler (95°C 1 min; 35 cycles of 94°C 30 sec, 58°C 30 sec, 72°C 1 min; 72°C 10 min; 4°C hold). Samples were stored (-20°C freezer).

The samples (5 μL product combined with 1.5 μL TrackIt Cyan/Orange 3X (Invitrogen)) were size fractionated at 100V for 30 min using 2% TBE with GelGreen agarose gel electrophoresis (RunOne Electrophoresis System, EmbiTech, San Diego California USA) and compared against a 100 bp DNA standard. The samples viewed for size aberrations (PrepOne Sapphire2 Imager, EmbiTech) and compared to the size of the relevant reference sequence fragment. For raw data see supplementary information figures 1-14.

Table 2 Primer sequences for fragment amplification in *MLH1*, *MSH2* and *MSH6* to test for deep intronic variants.

Oligo Name	Sequence (5' to 3')
MLH1 P1 FOR	AAGTTATCCAGCGGCCAGC
MLH1 P1 REV	CTTGCTCTGTATGCACACTTTCC
MLH1 P1 alt FOR	GTTCCCTGACGTGCCAGTC
MLH1 P1 REV	CTTGCTCTGTATGCACACTTTCC
MLH1 P2 FOR	CAGCATAAGCCATGTGGCTC
MLH1 P2 REV	CTGAATACCTGCCAACAACTTCC
MLH1 P3 FOR	CATGTGCTGGCAATCAAGG
MLH1 P3 REV	CATCCTGGAGGAATTGGAGC
MLH1 P4 FOR	ACATCGAGAGCAAGCTCCTG
MLH1 P4 REV	CTCCCGGAGAACCTCATGTC
MLH1 P5 FOR	CCGAAAGGAAATGACTGCAGC
MLH1 P5 REV	CACTTCACTCTGCTGGCCTGAG
MSH2 P1 FOR	CCAGGGGGTGATCAAGTACA
MSH2 P1 REV	TGAAACTGCAACCTGATTCTCC
MSH2 P2 FOR	TTGAAAGGCAAAAAGGGAGA
MSH2 P2 REV	CATGGTTTTCCACCTGATCC
MSH2 P3 FOR	CAGGCTCTGGAAAAACATGA
MSH2 P3 REV	GGGCCAGTAATGATGTGGAAC
MSH2 P4 FOR	GCCATTTTGGAGAAAGGACA
MSH2 P4 REV	CACCTTGCTCTCTTTCCAGATAG
MSH6 P1 FOR	CCAAGGCGAAGAACCTCA
MSH6 P1 REV	GTGCCTACCTCCATCTCTTCTTC
MSH6 P2 FOR	ACCAAGAAGGGCTGTAAACG
MSH6 P2 REV	CTTTCACCTGACATTATTCTGTCTG
MSH6 P3 FOR	TACGTCCCTGCTGAAGTGTG
MSH6 P3 REV	GCAAACTTCCCGAAATAATCG

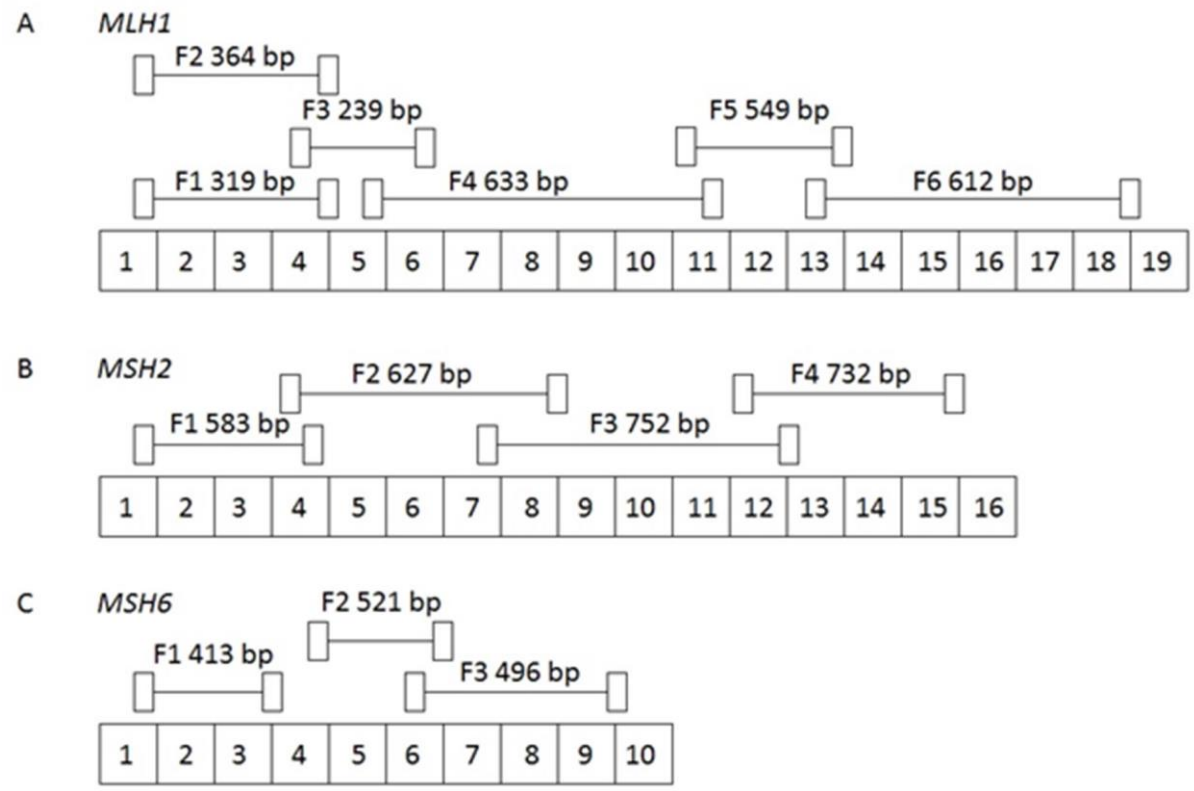


Figure 1 Location of the 13 fragments encompassing exon-exon boundaries of *MLH1*, *MSH2* and *MSH6*. Note (A) the 6 fragments for *MLH1* ranging from 239 bp up to 633 bp, F2 provides coverage for the alternative splice variant for exon 1 (B) the 4 fragments for *MSH2* ranging from 583 bp up to 752 bp, and (c) the 3 fragments for *MSH6* ranging from 413 bp up to 521 bp. All primer sets, where possible include one primer that spans an exon-exon junction to ensure no genomic DNA is amplified. Also, all primer sets overlap the previous to ensure complete coverage, except for exon 4 of *MSH6* which was too large.

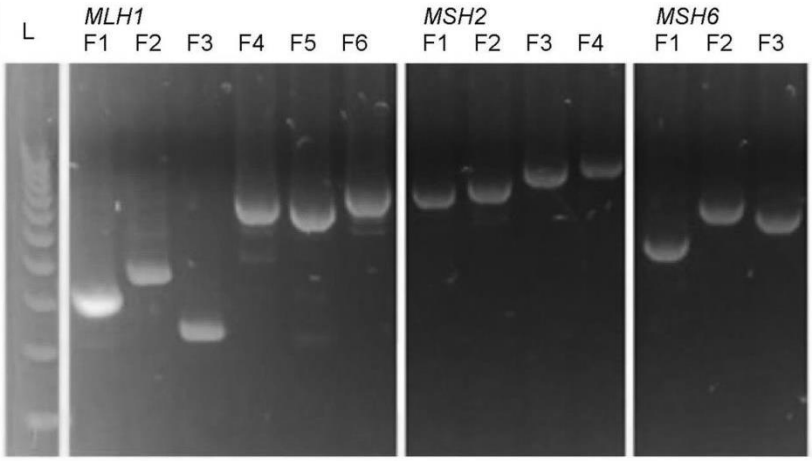


Figure 2 Primer optimization of the 13 fragments encompassing exon-exon boundaries of *MLH1*, *MSH2* and *MSH6.*

Validation of aberrant fragments

Fragments that were judged to be larger or smaller than the reference transcript were analysed by Sanger sequencing using a semi-automated DNA sequencer according to the manufacturers protocols (Applied Biosystems from the remainder of the amplified sample (not run on the agarose gel). Sequencing data was analysed using Mutation Surveyor (v3.0) and mutation pathogenicity was assessed using the LOVD database with reference to the Consensus InSight classification of variants[28].

Results and Discussion

This study aimed to reveal the presence of aberrant mRNA transcripts that were a result of large variants located deep within the introns of *MLH1*, *MSH2* or *MSH6*.

A total of thirteen fragments spanning the entire transcribed regions of *MLH1*, *MSH2* and *MSH6* from all samples derived from the forty-six HNPCC MMR mutation negative patients were successfully amplified and visualized by gel electrophoresis.

An unexpected observation was the lack of normal splice variants detected among the patient cohort whereas splice variants particularly in fragment four of *MLH1* were observed in all of the five control samples. The splice variants were confirmed by Sanger sequencing (see figure 3) and were associated with the skipping of exons nine and ten. Three other splice variants were identified in fragment five of *MLH*1 and fragments two and four of *MSH2* that were consistent with previously observed exon 12 splice variants in *MLH1* and exon 5 and exon 13 splice variants in *MSH2*.

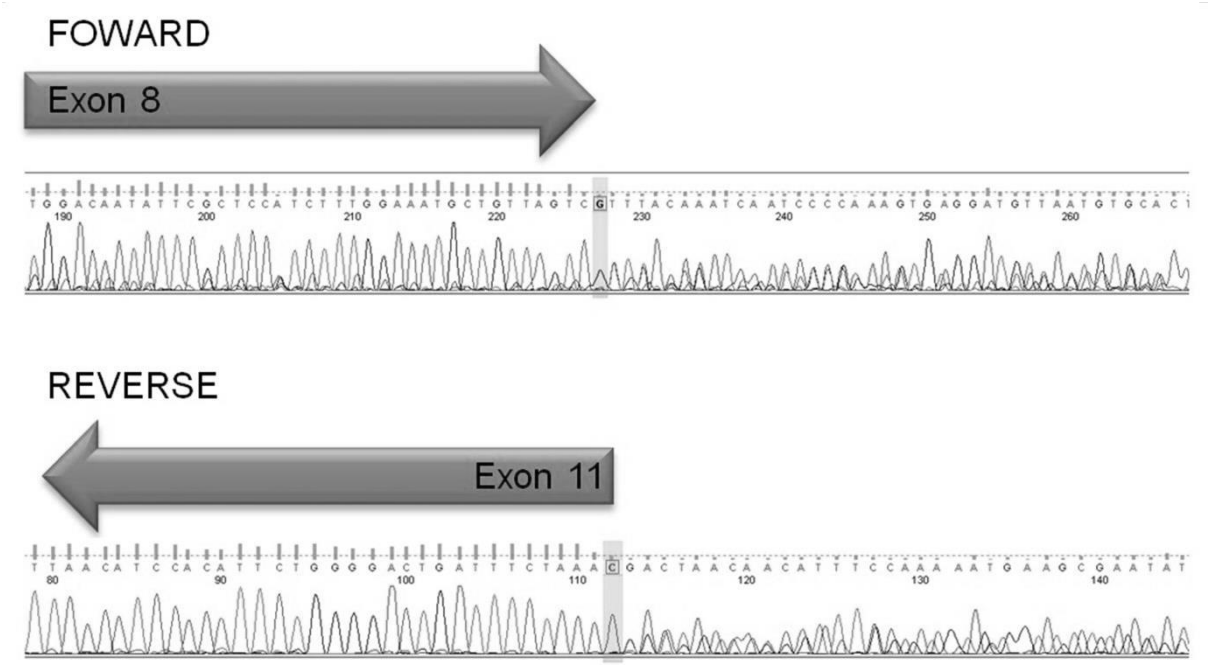


Figure 3 Example of a detected splice variant in a control showing the loss of exons 9 and 10 from *MLH1*. Loss indicated by frameshift pattern observed in forward primer immediately after the conclusion of exon 8 (highlighted last bp, G) and reverse primer immediately prior the start of exon 11 (highlighted first bp, C).

The absence of splice variants observed in the patients may be due to all the control RNA being derived from whole blood compared to patient RNA coming from EBV-transformed LCLs. It is recognised that EBV transformation can result in global changes in cellular isoform usage which may account for the absence of any evidence of differential splicing in the transformed patient samples[29]. This does not imply that cryptic splice sites deep within the intronic regions of the MMR genes under study would fail be in active.

With respect to our findings, we could not identify any aberrant transcripts that were the results of cryptic intronic variants in any of the cell lines studied. Several subtly different fragment sizes compared to the respective reference sizes were observed for eight of the fragments across the three genes in sixteen cell lines.

The PCR products of the suspected aberrantly sized fragments were successfully sequenced and aligned to the respective reference sequence; however no intronic or pathogenic variants were identified (see figure 4). A total of two polymorphisms were identified in two separate patient cell lines that resided in the flanking exonic sequence of the respective amplified fragments. The first cell line harboured a base substitution in fragment four of the *MLH1* gene which was shown to be in exon eight, c.655A>G. No sequence variants were identified in the initial sequencing of the genomic DNA from this patient and it is therefore thought to have arisen *de novo* as a result of EBV transformation[30]. The second cell line harboured a base substitution in fragment one of the *MSH2* gene that was in exon three, c.380A>G. Both the variants identified were considered to be non-pathogenic polymorphisms according to the InSight classification scheme[28]. As no frame-shifts were observed in the sequencing results this further supports the likely benign nature of these changes identified in this study. As these variants were considered non-deleterious and were not followed up further and nor was genomic DNA sequenced to confirm their presence.

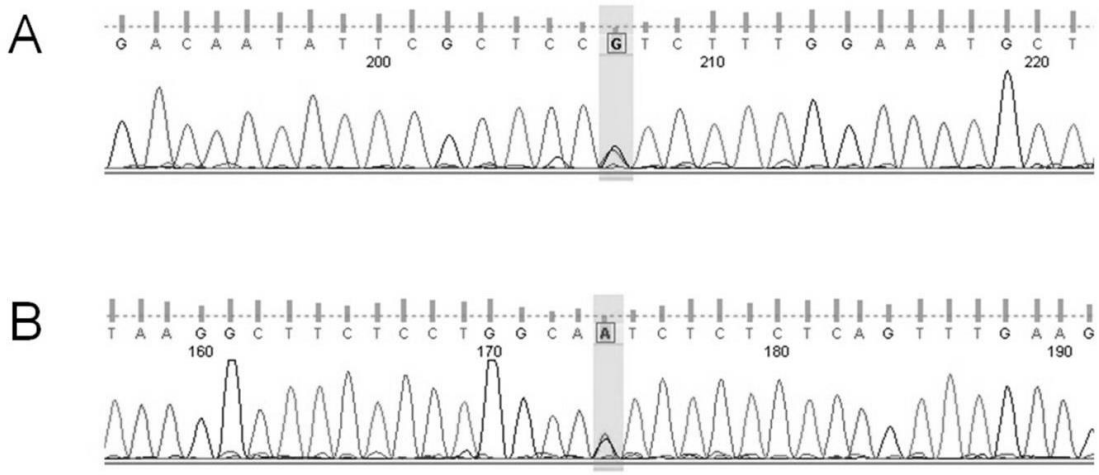


Figure 4 Sequencing results for the modest band changes. (A) a base substitution in fragment four of the *MLH1* gene which was shown to be in exon eight. c.655A>G (double peak present, lighter A and darker G), the called G peak being the mutation that is present, and (B) a base substitution in fragment one of the *MSH2* gene that was in exon three. c.380A>G (double peak present, lighter A and darker G), while called as A, the G peak is the mutation that is present. Location numbering relevant to base position within fragment, not position within gene. Both the variants identified were considered to be InSight Class 1, non-pathogenic polymorphisms[26].

Unlike in the previously reported FAP study where intronic variants resulting in aberrant splicing account for up to 8% (10 in 125 patients) of *APC* mutation negative cases, we found none in our investigation of forty-six HNPCC MMR mutation negative cases suggesting that they are far less common in this syndrome. Similarly, different frequencies of *de novo* mutations arising in HNPCC and FAP have been reported, where it is estimated *de novo* mutations occur in somewhere between 1% and 5% of HNPCC patients compared to 11% and 25% of FAP patients[31-33]. The observed difference between the frequency of transcript variants in FAP and HNPCC is significant (χ^2=3.9086, *p=0.04804*).

Conclusions

Overall the data presented in this study suggests that intronic insertions or deletions are rare events in HNPCC whereas they are relatively common in FAP. This might be related to the *de novo* mutation rate in the APC gene compared to that observed in the DNA MMR genes. Furthermore, the lack of normal splice variants observed in the HNPCC patients remains unclear and further studies are needed in larger patient populations to determine if they play a role in CRC.

List of Abbreviations

bp	base pair
C	celcius
CRC	colorectal cancer
EBV	Epstein- Barr virus
FAP	Familial adenomatous polyposis
FBS	Foetal Bovine Serum
HNEHREC	Hunter New England Human Research Ethics Committee
HNPCC	Hereditary non-polyposis colorectal cancer
HREC	University of Newcastle's Human Research Ethics Committee
LS	Lynch Syndrome
MLPA	Multiplex Ligation-dependant Probe Amplification
mM	micro molar
MMR	Mismatch Repair
MSI	Microsatellite instability
ng	nano grams
PCR	Polymerase Chain Reaction
RT	room temperature
RT-PCR	Real Time Polymerase Chain Reaction
uL	microliters
v	version

Competing Interests

The authors declare that they have no competing interests.

Author Contributions

ALM conducted the experiments, performed data analysis/interpretation and wrote the first draft of the manuscript.

BAT-P, T-JE and MW-B provided technical assistance.

TC and MAT assisted in the culturing the cell lines.

GNH provided critical review of the manuscript and helped design the experiments.

RJS conceived the study, designed the experimental approach and reviewed and approved the final version of the manuscript prior to submission.

Acknowledgements

This work has been supported by the following funding bodies and institutions: Australian Rotary Health/Rotary District 9650, the Commonwealth Scientific and Industrial Research Organization (CSIRO), the University of Newcastle and the Hunter Medical Research Institute. Samples were provided by the Hunter Area Pathology Service and the Hunter Community Study.

References

1. Rodriguez-Bigas, M.A. *et al.* A National Cancer Institute Workshop on Hereditary Nonpolyposis Colorectal Cancer Syndrome: meeting highlights and Bethesda guidelines. *J Natl Cancer Inst* **89**, 1758-62 (1997).

2. Vasen, H.F., Mecklin, J.P., Khan, P.M. & Lynch, H.T. The International Collaborative Group on Hereditary Non-Polyposis Colorectal Cancer (ICG-HNPCC). *Dis Colon Rectum* **34**, 424-5 (1991).

3. Kemp, Z., Thirlwell, C., Sieber, O., Silver, A. & Tomlinson, I. An update on the genetics of colorectal cancer. *Hum Mol Genet* **13 Spec No 2**, R177-85 (2004).

4. Peltomaki, P. Deficient DNA mismatch repair: a common etiologic factor for colon cancer. *Hum Mol Genet* **10**, 735-40 (2001).

5. Thompson, E. *et al.* Hereditary non-polyposis colorectal cancer and the role of hPMS2 and hEXO1 mutations. *Clin Genet* **65**, 215-25 (2004).

6. Fishel, R. *et al.* The human mutator gene homolog MSH2 and its association with hereditary nonpolyposis colon cancer. *Cell* **75**, 1027-38 (1993).

7. Lynch, H.T. & de la Chapelle, A. Hereditary colorectal cancer. *N Engl J Med* **348**, 919-32 (2003).

8. Parsons, R. *et al.* Hypermutability and mismatch repair deficiency in RER+ tumor cells. *Cell* **75**, 1227-36 (1993).

9. Vasen, H.F. & Boland, C.R. Progress in genetic testing, classification, and identification of Lynch syndrome. *JAMA* **293**, 2028-30 (2005).

10. Kuiper, R.P. *et al.* Recurrence and variability of germline EPCAM deletions in Lynch syndrome. *Hum Mutat* **32**, 407-14 (2011).

11. Ligtenberg, M.J. *et al.* 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 (2009).

12. Charbonnier, F. *et al.* Alternative splicing of MLH1 messenger RNA in human normal cells. *Cancer Res* **55**, 1839-41 (1995).

13. Clarke, L.A., Jordan, P. & Boavida, M.G. Cell type specificity in alternative splicing of the human mismatch repair gene hMSH2. *Eur J Hum Genet* **8**, 347-52 (2000).

14. Genuardi, M. *et al.* Characterization of MLH1 and MSH2 alternative splicing and its relevance to molecular testing of colorectal cancer susceptibility. *Hum Genet* **102**, 15-20 (1998).

15. Mori, Y. *et al.* Alternative splicing of hMSH2 in normal human tissues. *Hum Genet* **99**, 590-5 (1997).

16. Thompson, B.A., Martins, A. & Spurdle, A.B. A review of mismatch repair gene transcripts: issues for interpretation of mRNA splicing assays. *Clin Genet* **87**, 100-8 (2015).

17. Senter, L. *et al.* The clinical phenotype of Lynch syndrome due to germ-line PMS2 mutations. *Gastroenterology* **135**, 419-28 (2008).

18. McPhillips, M., Meldrum, C.J., Creegan, R., Edkins, E. & Scott, R.J. Deletion Mutations in an Australian Series of HNPCC Patients. *Hered Cancer Clin Pract* **3**, 43-7 (2005).

19. Bonis, P.A. *et al.* Hereditary nonpolyposis colorectal cancer: diagnostic strategies and their implications. *Evid Rep Technol Assess (Full Rep)*, 1-180 (2007).

20. Obermair, A. *et al.* Risk of Endometrial Cancer for women diagnosed with HNPCC-related colorectal cancer. *International journal of cancer* **127**, 7 (2010).

21. Petersen, S.M. *et al.* Functional examination of MLH1, MSH2, and MSH6 intronic mutations identified in Danish colorectal cancer patients. *BMC Med Genet* **14**, 103 (2013).

22. Clendenning, M. *et al.* Mutation deep within an intron of MSH2 causes Lynch syndrome. *Fam Cancer* **10**, 297-301 (2011).

23. Spier, I. *et al.* Deep intronic APC mutations explain a substantial proportion of patients with familial or early-onset adenomatous polyposis. *Hum Mutat* **33**, 1045-50 (2012).

24. Aretz, S. *et al.* Familial adenomatous polyposis: aberrant splicing due to missense or silent mutations in the APC gene. *Hum Mutat* **24**, 370-80 (2004).

25. Kaufmann, A. *et al.* Analysis of rare APC variants at the mRNA level: six pathogenic mutations and literature review. *J Mol Diagn* **11**, 131-9 (2009).

26. Auclair, J. *et al.* Systematic mRNA analysis for the effect of MLH1 and MSH2 missense and silent mutations on aberrant splicing. *Hum Mutat* **27**, 145-54 (2006).

27. Vasen, H.F., Watson, P., Mecklin, J.P. & Lynch, H.T. New clinical criteria for hereditary nonpolyposis colorectal cancer (HNPCC, Lynch syndrome) proposed by the International Collaborative group on HNPCC. *Gastroenterology* **116**, 1453-6 (1999).

28. Thompson, B.A. *et al.* Application of a 5-tiered scheme for standardized classification of 2,360 unique mismatch repair gene variants in the InSiGHT locus-specific database. *Nat Genet* **46**, 107-15 (2014).

29. Homa, N.J. *et al.* Epstein-Barr virus induces global changes in cellular mRNA isoform usage that are important for the maintenance of latency. *J Virol* **87**, 12291-301 (2013).

30. Londin, E.R. *et al.* Whole-exome sequencing of DNA from peripheral blood mononuclear cells (PBMC) and EBV-transformed lymphocytes from the same donor. *BMC Genomics* **12**, 464 (2011).

31. Win, A.K. *et al.* Determining the frequency of de novo germline mutations in DNA mismatch repair genes. *J Med Genet* **48**, 530-4 (2011).

32. Bisgaard, M.L., Fenger, K., Bulow, S., Niebuhr, E. & Mohr, J. Familial adenomatous polyposis (FAP): frequency, penetrance, and mutation rate. *Hum Mutat* **3**, 121-5 (1994).

33. Ripa, R., Bisgaard, M.L., Bulow, S. & Nielsen, F.C. De novo mutations in familial adenomatous polyposis (FAP). *Eur J Hum Genet* **10**, 631-7 (2002).

CHAPTER 5: COPY NUMBER VARIATION IN HEREDITARY BREAST CANCER

Introduction

Global cancer statistics identifies breast cancer as the most frequently diagnosed cancer and leading cause of cancer related death in females[19]. Nearly 27% of breast cancers arise in a familial setting, characteristically displaying an earlier age of disease diagnosis and a higher frequency among family members[217,218].

Currently *BRCA1* and *BRCA2* represent the most frequently tested breast cancer susceptibility genes; however they account for less than 20% of all hereditary breast cancer patients[205,354,355]. For many patients coming from breast cancer families no genetic diagnosis can be found. This suggests that either other genes are involved in disease risk or other genomic events may result in the inactivation of known breast cancer susceptibility genes.

Copy number variants (CNVs), or regions of duplication or deletion in the genome are considered likely contributors to disease as a consequence of them often encompassing large stretches of genomic sequence[278,343-348] and are yet to be explored in a large cohort of *BRCA1*/*BRCA2* mutation negative hereditary breast cancer patients. Two reports have recently examined CNVs in association with *BRCA1* and *BRCA2* mutation negative hereditary breast cancer patients. The first of these reported a greater abundance of rare CNVs in breast cancer patients and suggest that rare CNVs are likely to contain genetic factors associated with disease predisposition, while the second report suggested several CNV markers likely to be associated with familial breast cancer risk which may be useful for assessment of disease risk[285,290].

This part of the thesis aims to describe the CNV landscape in *BRCA1* and/or *BRCA2* mutation negative hereditary breast cancer patients and conduct a patient-control analysis to identify CNVs which could be associated with the genetic basis of their disease.

Publication

Amy L. Masson, Bente A. Talseth-Palmer, Tiffany-Jane Evans, Desma M. Grice, Garry N. Hannan and Rodney J. Scott (2014) Expanding the genetic basis of copy number variation in familial breast cancer, *Hereditary Cancer in Clinical Practice*, 12:15.

Co-author statement

I attest that Research Higher Degree candidate Amy Louise Masson contributed to the above manuscript including involvement in the conception and design of the study; conducting the laboratory work, data analysis and interpretation; and preparation of the manuscript.

Co-author	Signature	Date
Bente A. Talseth-Palmer		
Tiffany-Jane Evans		
Desma M. Grice		
Garry N. Hannan		
Rodney J. Scott		

Amy Louise Masson Date: 01/09/2015

Professor Robert Callister Date: 01/09/2015

Assistant Dean Research Training

Expanding the genetic basis of copy number variation in familial breast cancer

Amy L. Masson[1,4], Bente A. Talseth-Palmer[1,4], Tiffany-Jane Evans[1,4], Desma M. Grice[1,2], Garry N. Hannan[2] and Rodney J. Scott[1,3,4,*]

[1] Information Based Medicine Program, Hunter Medical Research Institute, University of Newcastle, Newcastle, New South Wales, 2305 Australia

[2] CSIRO Preventative Health Flagship and Animal, CSIRO Food and Health Sciences Division, North Ryde, New South Wales, 2113 Australia

[3] Division of Molecular Medicine, Hunter Area Pathology Service, John Hunter Hospital, Newcastle, New South Wales, 2305 Australia

[4] School of Biomedical Sciences and Pharmacy, Faculty of Health, University of Newcastle, New South Wales, 2308 Australia

*Author to whom correspondence should be addressed; E-mail: Rodney.Scott@newcastle.edu.au (R.J.S.);

Tel.: +61 (2) 4921 4974;

Fax: +61 (2) 4921 4253

Abstract

Introduction

Familial breast cancer (fBC) is generally associated with an early age of diagnosis and a higher frequency of disease among family members. Over the past two decades a number of genes have been identified that are unequivocally associated with breast cancer (BC) risk but there remain a significant proportion of families that cannot be accounted for by these genes. Copy number variants (CNVs) are a form of genetic variation yet to be fully explored for their contribution to fBC. CNVs exert their effects by either being associated with whole or partial gene deletions or duplications and by interrupting epigenetic patterning thereby contributing to disease development. CNV analysis can also be used to identify new genes and loci which may be associated with disease risk.

Methods

The Affymetrix Cytogenetic Whole Genome 2.7M (Cyto2.7M) arrays were used to detect regions of genomic re-arrangement in a cohort of 129 fBC *BRCA1*/*BRCA2* mutation negative patients with a young age of diagnosis (<50 years) compared to 40 unaffected healthy controls (>55 years of age).

Results

CNV analysis revealed the presence of 275 unique rearrangements that were not present in the control population suggestive of their involvement in BC risk. Several CNVs were found that have been previously reported as BC susceptibility genes. This included CNVs in *RPA3*, *NBN* (*NBS1*), *MRE11A* and *CYP19A1* in five unrelated fBC patients suggesting that these genes are involved in BC initiation and/or progression. Of special interest was the identification of *WWOX* and *FHIT* rearrangements in three unrelated fBC patients.

Conclusions

This study has identified a number of CNVs that potentially contribute to BC initiation and/or progression. The identification of CNVs that are associated with known tumour suppressor genes is of special interest that warrants further larger studies to understand their precise role in fBC.

Key Words

Breast Cancer, DNA repair, CNV

Introduction

Global cancer statistics identify BC as the most frequently diagnosed cancer (23%) and leading cause of cancer related death (14%) in females[1]. Nearly 27% of these BCs occur in a familial setting typically associated with an earlier age of disease diagnosis and a higher frequency among family members and is termed fBC[2,3]. It is estimated that 5-10% of these families harbor germline mutations or complex genomic changes that render inactive one of four high penetrance genes (*BRCA1*, *BRCA2*, *TP53* or *PTEN*) or moderate penetrance genes (*CHEK2*, *ATM*, *BRIP1* and *PALB2*)[2,4,5]. Associations have also been identified for other genes in fBC including *ATM*, *CASP8*, *CTLA4*, *NBN*, *CYP19A1*, *TERT* and *XRCC3*[6]**.** The most recent BC meta-analysis has identified 41 loci and suggests that over 1000 loci may be involved in disease susceptibility[7]. The identification of *BRCA1* and *BRCA2* as susceptibility genes for BC and the more recent addition of *PALB2*, *BRIP1* and *RAD51C*[5] have focused attention on genes associated with double strand break repair (DSBR). There are at least 39 genes implicated in DSBR, all of which could potentially be associated with BC risk. This is analogous to DNA mismatch repair (MMR), where there are at least 21 genes associated with this process, of which four are now routinely assessed and more recently a fifth gene (*POLD1*) has been added to the list[8,9]. Despite the plethora of information regarding genetic loci associated with BC risk, for many fBC cases no genetic predisposition has been identified. Outside the context of gene mutations other mechanisms may be associated with disease development including gene silencing as a result of epigenetic re-programming of BC susceptibility genes (analogous to loss of *EPCAM* and the re-arrangement of the epigenetic profile on chromosome 2, rendering *MSH2* inactive[10,11]), or mutations in genes not yet associated with a predisposition to disease.

One type of genetic alteration that could account for susceptibility is genetic re-arrangements detected as CNVs. CNVs represent a class of structural variation involving regions of duplication or deletion of genomic material that can encompass large stretches of genomic sequence ranging from megabases (Mbs) to a few kilobases (Kb) in size. As a consequence, CNVs can contribute to disease when they incorporate functional gene sequence (coding and promoter regions of genes) or exert more cryptic effects,that could affect epigenetic regulation (methylation, microRNA targets) and non-coding intronic gene sequences[12-23]. Two reports have recently examined CNVs in association with *BRCA1*/*BRCA2* mutation negative fBC patients. The first of these has reported a greater abundance of rare CNVs in fBC patients and

suggest that rare CNVs are likely to contain genetic factors associated with BC predisposition, while the second report associated several CNV markers with fBC risk and suggests their use in disease risk assessment[24,25].

The detection of CNVs has historically relied upon the use of DNA arrays, typically comprised of oligonucleotide markers distributed across the whole genome. The resolution of DNA arrays has increased to allow for the detection of genomic rearrangements as small as a few Kb in size. In this study we used the Affymetrix Cyto2.7M array which provided the highest genomic coverage of any commercially available microarray at the time of assay to assess CNV variation in an fBC cohort. The Cyto2.7M array contains a combination of 400,000 single nucleotide polymorphisms (SNPs) and >2.1 million copy number probes (average spacing 1395 base pairs (bp)) which together can be used to accurately detect genomic rearrangements.

We conducted a patient-control analysis examining 129 fBC patients and 40 control subjects derived from the same population to identify CNVs which could be associated with the genetic basis of their disease. To date this study represents one of the largest CNV studies of *BRCA1*/*BRCA2* mutation negative fBC patients

Materials and Methods

Samples

The study was approved by the University of Newcastle's Human Research Ethics Committee and the Hunter New England Human Research Ethics Committee. Genomic DNAs were obtained from fBC patients who had given informed consent for their DNA to be used for studies into their disease and control DNA samples from the Hunter Community Study (HCS)[26]. DNA was extracted from whole blood by the salt precipitation method[27].

Cohorts of 129 patients clinically diagnosed with early-onset fBC were used in this study. All patients had been diagnosed with BC and were the first individual (proband) of their family to seek genetic testing for mutations in *BRCA1*/*BRCA2*. Mutation screening was performed using Sanger Sequencing and Multiplex ligation-dependant probe amplification (MLPA) analysis. No mutations were identified in any of the patients (*BRCA1*/*BRCA2* mutation negative). Average age of patients was calculated to be <40.7 years. Genomic DNA from 40 controls[26] was also utilized in this study. These were healthy (cancer free) individuals aged >55 years at the time of sample collection.

Genomic array preparation and data processing

The genomic DNA from 129 fBC patients and 40 controls were processed on the Affymetrix Cyto2.7M array consistent with manufacturer's protocols. CEL files were analysed in Affymetrix, the Chromosome Analysis Suite (ChAS) (Version CytoB-N1.2.0.232; r4280) using NetAffx Build 30.2 (Hg18) annotation. Quality control (QC) parameters were optimized and validated using a training set of 20 randomly selected samples. All samples were subject to a series of quality cut-off measures: snpQC >1.1 (SNP probe QC based off distances between the distribution of alleles (AA, AB and BB) where larger differences are associated with an increased ability to differentiate genotype; default), mapdQC <0.27 (Median Absolute Pair-wise Difference; CN probe QC based off a reference model; default) and wavinessSd <0.1 (measure of standard deviation in data waviness; the GC content across the genome correlates with average probe intensities i.e. high GC probes are brighter than low GC probes on average, creating waves in the data). CNV regions were assessed according to call confidence, probe count, size and by visual inspection for distinction from normal CN state. Data was also visually inspected to identify regions with low density of markers (supplementary table 1) which were excluded across all samples. Most thresholds

were more stringent than default settings alone in an aim to minimize false-positive CNVs being included in the analysis. CNV regions were filtered across all samples using the following parameters: >90% confidence, autosomes only and a minimum number of 24 probes. Using these parameters the limit of detection was 9.65 Kb across all samples used in the current study. This does not exclude the possibility of CNVs smaller than this from contributing to disease in a proportion of fBC patients.

CNV and statistical analysis

CNVs in fBC patients and controls were subject to a series of comprehensive analyses which included: (1) interrogation for CNVs residing in or ±100 Kb of 61 genes (associated with DSBR, MMR and BC susceptibility) and 41 SNPs recently reported to be associated with BC risk[6,7,28,29] (see supplementary tables 2 and 3); (2) comparison of CNVs between fBC patients and controls according to CN occurrence and distribution across the genome; (3) identification of rare CNVs using the Database of Genomic Variants (DGV); and (4) the identification of genes associated with malignancy (non-specific) using the Network of Cancer Genes (Version 3.0) and the Cancer Gene Census (CGC; 15 March 2012) databases[30,31]. Associations (e.g. numbers and sizes of CNVs) were statistically compared using a two tailed un-paired t-test Graphpad Prism (Version 6)[32].

Validation of CNV results

CNV results were validated using pre-designed TaqMan Copy Number (CN) Assays (Applied Biosystems). Up to two CN assays were selected within the CNV region indicated by the Cyto2.7M array and CN assays, proximal but external to the region were also selected as controls (assay information summarized in supplementary table 4). A total of 11 samples were run in triplicate comprised of the sample(s) of interest, a calibrator (control) sample with known CN for the region of interest and a no-template-control (NTC). Real-time PCR was conducted according to manufacturer's protocols using 10 ng of DNA sample in a final reaction volume of 20 μL. The assay was run on the real-time PCR machine (Applied Biosystems 7500; SDS software Version v1.4) according manufacturer's protocols. The results were exported to CopyCaller v2.0 software (Applied Biosystems) for analysis.

Three CNVs were validated using this secondary independent assay (supplementary table 5). The CNVs included a CN gain and a CN loss in the *WWOX* gene as well as a CN loss in the *FHIT* gene. Given the high concordance between the CNV calling within the experimental parameters set for this study and the independent copy number

assays we considered that it was not necessary to confirm all CNVs using a second independent assay.

Results

Array resolution and CNV detection

Analysis of Cyto2.7M array data revealed a total of 414 CNVs in 169 individuals assessed in this study (table 1). CNVs detected ranged in size from 9.65 Kb to 1335.06 Kb. There was no difference in the average number of CNVs identified in the patients versus the controls (p=0.75). The average genomic burden of CNVs also did not differ between patients (226.93 Kb) and controls (295.52 Kb), p=*0.30*; or the average CNV size between patients (76.22 Kb) and controls (106.57 Kb), s, p=*0.07*.

Table 1 Summary of CNV results from the BC patients and control participants.

		CNV Count			CNV Size (Kb)		
		Total CNVs per group	Median CNVs per sample	Mean CNVs per sample	Total CNV affected genome per group	Mean total CNV affected genome per sample	Mean size of a CNV
Patients	129	310	2	2.40	29273.63	226.93	76.22
Controls	40	104	2	2.60	11820.75	295.52	106.57
p	-	-	-	*0.75*	-	*0.30*	*0.07*

Occurrence and distribution of CNVs in fBC patients

Overall 310 CNVs were identified in fBC patients of which 35 also occurred in controls (supplementary table 6). Since these regions were represented in the control population they were removed from further analysis. Of the 275 CNVs unique to the patients (supplementary table 7), 94 have was previously described in the DGV and 39 spanned genomic regions that were common to multiple patients (table 2). Of these 11 CNVs (located on chromosomes 2, 3, 4, 6, 11, 14, 15, 17 and 18) were common to two patients; three were common to three patients (located on chromosomes 4, 5 and 19); and two were common to four patients (located on chromosomes 3 and 18). Among these, three genomic regions (located chromosomes 6, 11 and 19) were considered novel (not reported in the DGV) and likely to represent regions of potential association with BC risk.

Of the CNVs unique to patients 160 (58.18%) encompassed genes. A CNV located in *SUPT3H* was also excluded from analysis as it was identified to be affected by a re-arrangement in a control sample and considered unlikely to be associated with disease risk. Therefore a total of 159 genes were associated with a CNV were identified as being unique to the fBC patients and represent genes potentially associated with disease. A total of 24 genes associated with 44 CNVs (gains, losses or both) were identified in multiple individuals (as shown in table 3): 19 genes, including *LAMB3*, *NBN*, *IL8* and *WWOX*, were affected by a CNV in two individuals; *PIK3R5* and *POU2F3* were affected by a CNV in three individuals; *ARHGEF12* and *TMEM136* were affected by a CNV in four individuals; and *NAMPT* was affected by a CNV in five individuals.

Table 2 Genomic regions associated with unique CNVs identified in multiple patients.

Type	Chr	Start (bp)*	End (bp)*	Size (Kb)	Probes
2 CNV gains					
Gain	2	13,119,088	13,199,687	80.6	48
Gain	2	13,135,013	13,199,687	64.7	43
Gain	2	82,055,473	82,163,764	108.3	85
Gain	2	82,056,404	82,168,370	112.0	89
Gain	3	958,296	1,012,953	54.7	33
Gain	3	975,908	1,032,700	56.8	29
Gain	6	27,738,385	27,764,062	25.7	26
Gain	6	27,742,403	27,770,374	28.0	24
Gain	15	79,783,294	79,876,946	93.7	77
Gain	15	79,795,446	79,876,343	80.9	70
Gain	17	21,503,478	21,648,413	144.9	25
Gain	17	21,503,478	21,650,626	147.2	26
3 CNV gains					
Gain	4	25,672,202	25,703,024	30.8	31
Gain	4	25,678,621	25,710,178	31.6	32
Gain	4	25,680,434	25,710,412	30.0	31
Gain	5	59,749,693	59,807,906	58.2	51
Gain	5	59,749,693	59,807,906	58.2	51
Gain	5	59,749,693	59,810,944	61.3	52
Gain	19	36,911,234	36,939,557	28.3	36
Gain	19	36,918,927	36,940,929	22.0	32
Gain	19	36,918,927	36,944,555	25.6	36
2 CNV losses					
Loss	11	95,844,428	95,917,476	73.1	54
Loss	11	95,844,428	95,917,476	73.1	54
Loss	14	44,229,915	44,294,996	65.1	53
Loss	14	44,229,915	44,294,996	65.1	53
Loss	17	19,439,549	19,476,055	36.5	28
Loss	17	19,439,549	19,476,055	36.5	28
Loss	18	1,714,779	1,828,901	114.1	109
Loss	18	1,714,779	1,828,901	114.1	109
4 CNV losses					

Loss	3	166,523,809	166,565,186	41.4	39
Loss	3	166,523,809	166,565,186	41.4	39
Loss	3	166,523,809	166,566,558	42.8	40
Loss	3	166,525,250	166,565,186	39.9	38
Loss	18	1,894,368	1,974,284	79.9	63
Loss	18	1,894,368	1,974,284	79.9	63
Loss	18	1,894,368	1,974,284	79.9	63
Loss	18	1,894,368	1,974,284	79.9	63
2 CNV gain and loss					
Gain	4	160,917,340	161,068,954	151.6	119
Loss	4	160,983,513	161,011,918	28.4	29

Probes = number of markers within a CNV segment

* set at first and last marker associated with the respective CNV

Table 3 Genes associated with unique CNVs identified across multiple patients.

CNV	Number of Patients	Gene	Loci
Gains	2	*B2M*	15q21.1
	2	*DSCAM*	21q22.2
	2	*G0S2*	1q32.2
	2	*GNG2*	14q22.1
	2	*GPR98*	5q14.3
	2	*IL8*	4q13.3
	2	*LAMB3*	1q32.2
	2	*LIMS1*	2q13
	2	*NBN*	8q21.3
	2	*TAGAP*	6q25.3
	2	*TRIM69*	15q21.1
Both	2	*CNTN4*	3p26.3
	2	*IMMP2L*	7q31.1
	2	*WWOX*	16q23.1
Losses	2	*ACYP2*	2p16.2
	2	*PCDH9*	13q21.32
	2	*SPINT4*	20q13.12
	2	*TSPYL6*	2p16.2
	2	*VAV3*	1p13.3
	3	*PIK3R5*	17p13.1
	3	*POU2F3*	11q23.3
	4	*ARHGEF12*	11q23.3
	4	*TMEM136*	11q23.3
	5	*NAMPT*	7q22.2

Rare CNVs in fBC patients

There were 95 rare CNVs identified in 42 of the fBC patients. Of these 70 were associated with 78 genes and were found in 27 patients. Out of the 78 genes *SUPT3H* was excluded from further analysis as it was identified in a healthy control subject. Ten genes that were disrupted due to the presence of a CNV had previously been associated with cancer[30,31] including *ARHGAP26*, *ARHGEF12*, *CARD11*, *CPD*, *FAM135B*, *TSHR*, *MLLT11*, *PTK2B*, *RHOH* and *FHIT* (table 4). The remaining CNVs affecting 67 genes were unique and have not previously been associated with malignancy (listed in supplementary table 8). These genes potentially represent new candidates that require further investigation.

Table 4 Results for the ten CNVs associated with seven patients which affect genes previously associated with cancer. Gene, age of patient diagnosis (Dx), CNV type (gain or loss), location (chromosome, start and end) and CNV size are indicated.

Genes	Dx	Type	Chr	Start (bp)	End (bp)	Size (Kb)
FHIT	22	Loss	3	60,494,885	60,632,282	137.4
CARD11	37	Gain	7	2,946,394	2,996,375	50
FAM135B	38	Gain	8	139,259,837	139,306,535	46.7
ARHGEF12	51	Gain	11	119,697,081	119,723,342	26.3
TSHR	~49	Gain	14	80,659,512	80,669,166	9.7
MLLT11	46	Gain	1	149,289,549	149307059	17.5
CPD		Gain	17	25,700,671	25,756,973	56.3
RHOH	28	Gain	4	39,864,888	39,888,181	23.3
ARHGAP26		Gain	5	142,147,309	142,174,652	27.3
PTK2B		Gain	8	27,237,115	27,333,842	96.7

Genomic changes involving BC susceptibility genes or the recently identified BC susceptibility loci

There are at least 61 genes including those involved in DNA DSBR and MMR that could potentially contribute to fBC[6,7,28,29]. CNV data for the 129 fBC patients and 40 controls was screened for genomic re-arrangements within or ±100 Kb either side of these 61 genes. Five patients were identified to harbour CN gains located within or in the vicinity of four genes (table 5): one within *RPA3* gene; two within the *NBN* gene; one 55.7 Kb upstream of the *MRE11A* gene and one other 89.2 Kb upstream of the *CYP19A1* gene. All gains are predicted to result in disruption of the respective genes' coding sequence (via the insertion of additional genomic material which is expected to result in loss of function). With respect to the *NBN* gene a CNV loss was also identified in a control residing in a region located 52.6 Kb downstream of the gene but did not appear to be associated with disruption of the coding sequence. No CNVs were identified that were located in the same 41 genomic regions that have recently been reported as BC susceptibility loci[7].

The identification of a CNV that involved *WWOX* in two unrelated patients (see table 6, figures 1 and 2) was of interest as this gene is located in a fragile site (*FRA16D*) associated with cancer development and has been shown to interact with *TP53* and *ACK1*[33] and has recently been reported to be involved in breast carcinogenesis[34,35]. Together, this suggests that loss of function of *WWOX* could potentially be involved in BC susceptibility. One patient harboured a CNV gain that was predicted to disrupt the coding sequence of the gene via the insertion of additional genomic material whereas the other patient had a CNV loss that is expected to result in loss of function. Both of these changes were confirmed using an independent CN assay (see supplementary table 5). A number of recent reports have also correlated BC development with changes in the *FHIT* gene which similarly to *WWOX* is located in a fragile site (*FRA3B*) and has again been linked to tumour development[36-43]. CNV analysis revealed a CN loss that encompassed *FHIT* (table 6 and figure 3) which was confirmed using an independent assay (supplementary table 5).

Table 5 Search results for regions containing CN gains and CN losses within ±100 Kb the 61 genes associated with BC risk. CNV location (chromosome, start bp and end bp), size (Kb) and type; as well as the gene affected by the variant are indicated.

Cohort	Genes	Type	Chr	Start (bp)	End (bp)	Size (Kb)
Patients	*RPA3*	Gain	7	7,670,435	7,697,631	27.2
	NBN	Gain	8	91,048,149	91,070,004	21.9
	NBN	Gain	8	91,050,795	91,088,236	37.4
	55.7 Kb upstream *MRE11A*	Gain	11	93,922,391	93,960,356	38.0
	89.2 Kb upstream *CYP19A1*	Gain	15	49,507,272	49,579,058	71.8
Control	52.6 Kb downstream *NBN*	Loss	8	90,913,791	90,962,106	48.3

Table 6 CNVs associated with fragile site *FRA16D* and *FRA3B*. CNV location (chromosome, start bp and end bp) and size (Kb); as well as the confidence score associated with CNV call, the gene affected by the variant, the number of probes used to call the CNV and if the variant has previously been reported in the DGV.

Chr	Start (bp)	End (bp)	Size (Kb)	Gene	Probes	DGV
16	76,684,338	76,929,109	244.8	*WWOX*	222	Reported
16	76,947,909	77,009,160	61.3	*WWOX*	69	Reported
3	60,494,885	60,632,282	137.4	*FHIT*	158	

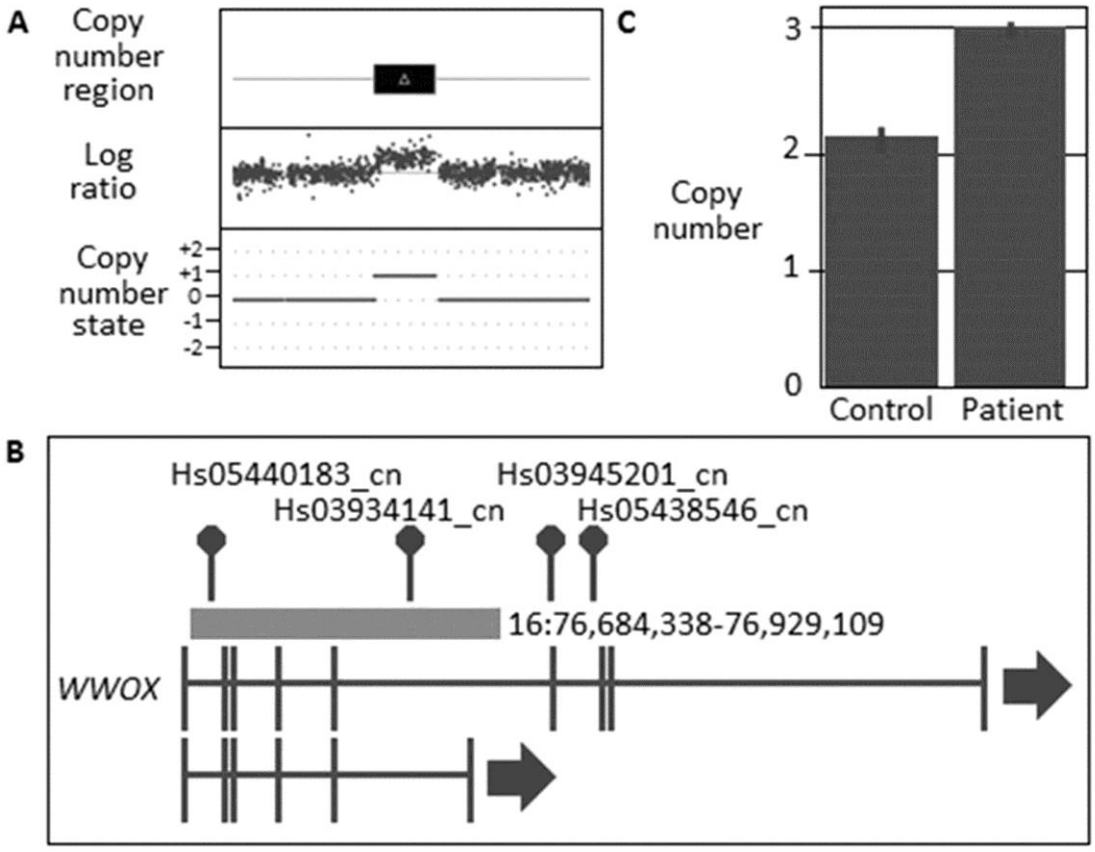


Figure 1 CNV results for *WWOX* duplication in fBC patient. (A) CNV profile from Cyto2.7M array data defining the region of duplication including the genomic state (where 0 = the normal two copies and +1 = one extra copy; (B) Location of the duplication within the gene and with respect to the CN assays used in validating the variant; and (C) TaqMan CN Validation assay showing the duplication represented by Hs03934141_cn: note the normal two copies of this region identified in the control, confirmation of the aberrant three copies in the fBC patient and the CN range bars associated with the three technical replicates used to validate the CNVs.

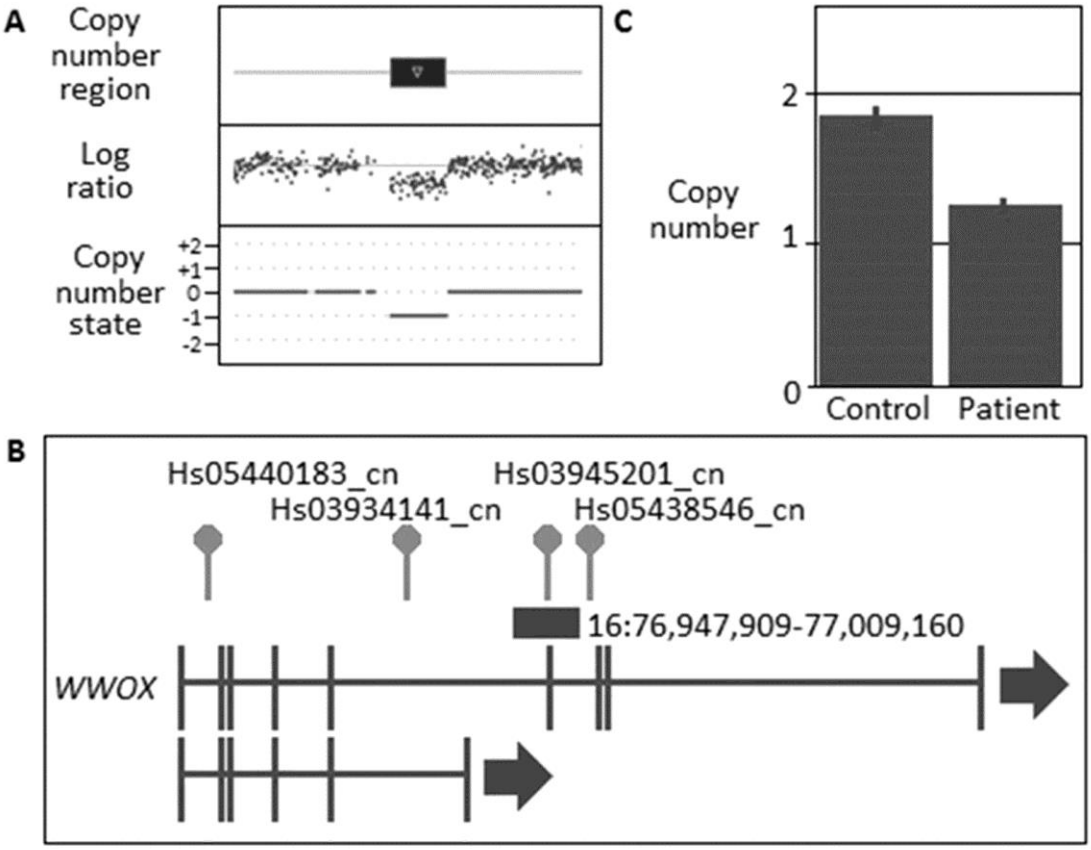


Figure 2 CNV results for *WWOX* deletion in fBC patient. (A) CNV profile from Cyto2.7M array data defining the region of deletion including the genomic state (where 0 = the normal two copies and -1 = one less copy; (B) Location of the deletion within the gene and with respect to the CN assays used in validating the variant; and (C) TaqMan CN Validation assay showing the deletion represented by Hs03945201_cn: note the normal two copies of this region identified in the control, confirmation of the aberrant one copy in the fBC patient and the CN range bars associated with the three technical replicates used to validate the CNVs.

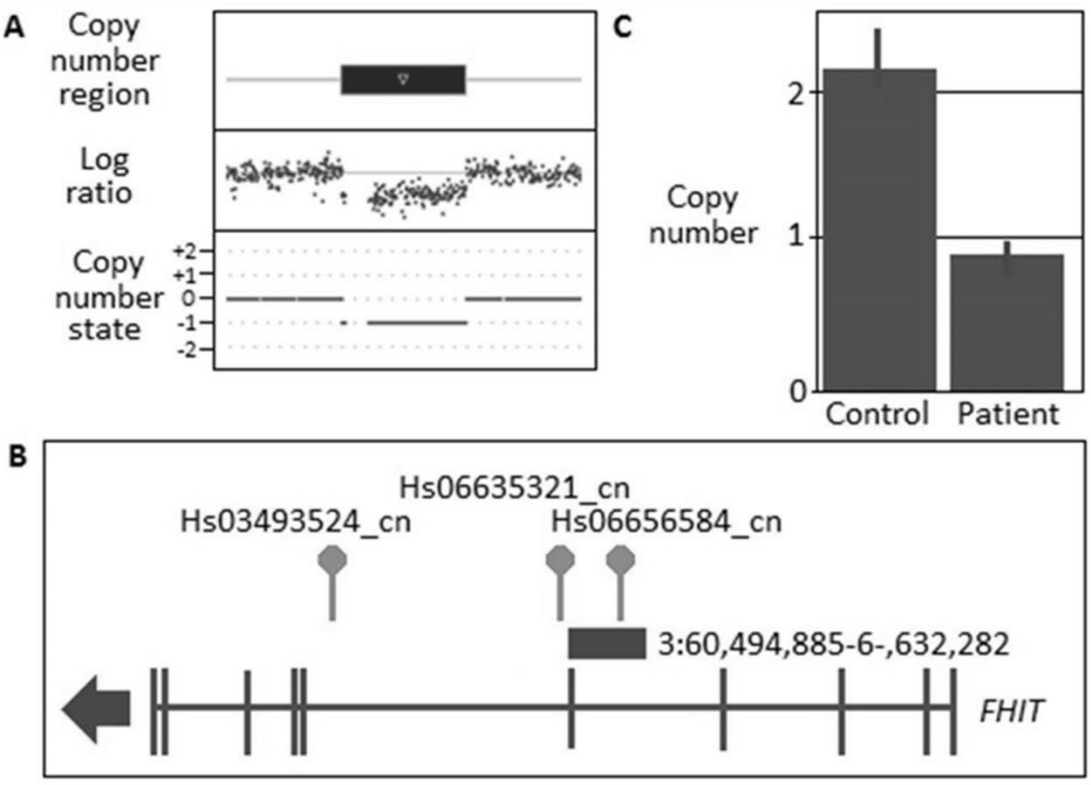


Figure 3 CNV results for *FHIT* deletion in fBC patient. (A) CNV profile from Cyto2.7M array data defining the region of deletion including the genomic state (where 0 = the normal two copies and -1 = one less copy; (B) Location of the deletion within the gene and with respect to the CN assays used in validating the variant; and (C) TaqMan CN Validation assay showing the deletion represented by Hs06656584_cn: note the normal two copies of this region identified in the control, confirmation of the aberrant one copy in the fBC patient and the CN range bars associated with the three technical replicates used to validate the CNVs.

Discussion

The association between CNVs and fBC is yet to be fully defined. In this study we provide evidence that CNVs are a potential explanation for small but significant number of fBC patients who do not harbour germline mutations in known susceptibility genes.

Genomic resolution provided by microarray technology has increased significantly allowing for the discovery of ever smaller CNVs. The resolution of the array used in this study was limited to the identification of CNVs greater than 9.65 Kb in size, and hence we cannot rule out the potential involvement of smaller CNVs in the aetiology of fBC. There have been a number of technical issues associated with the identification of CNVs that have compounded the difficulties in assessing the role of genomic rearrangements in disease. Different array platforms, software algorithms, batch effects and population stratification influence the accuracy of calls made to and comparisons of CNV data[44-46]. To help in reducing the influence of these effects a set of 40 older population controls was used as the basis to differentiate between CNVs associated with breast cancer and uninformative controls. All samples (both cases and controls) were processed on one platform and analysed using the same analysis software and experimental parameters. Comparison between the number and size of CNVs between patients and controls did not reveal any significant differences between cohorts. It is important to note the limited number of controls utilized in the current study represents a potential bias, however it is reassuring to note that despite this potential limitation, our observations are consistent with two previous reports on fBC (68 patients and 100 controls) and *BRCA1*-associated ovarian cancer (84 patients and 47 controls)[24,47].

We also identified 67 genes associated with novel CNVs that have yet to be linked with BC risk. It is interesting to note that many of these have been implicated in biological processes involving metabolism and biological regulation[48]. This provides the basis for further investigation into expanding the number of genes involved in BC development.

Our study has identified CNVs in close proximity to a number of genes previously associated with BC risk in a fBC cohort: *ARHGEF12* has been proposed to be a candidate tumour suppressor gene in BC whereby its under expression (typically as a result of genomic loss) has been observed in BC cell lines and where re-induction of the gene resulted in reduced cell proliferation and colony formation[49]; Laminin 5 (LN5) genes (including *LAMB3*) have been shown to exhibit reduced expression as a result of epigenetic inactivation in 65% of BC cell lines[50]; *NBN* has been recently reported to be

associated with BC risk[6]; and *NAMPT* has been shown to modify the effects of *PARP* inhibitors used in the treatment of triple-negative BCs suggesting the potential for a combination of *NAMPT* and *PARP* inhibitors in the treatment of this disease[51].

Of all the genes affected by a CNV identified in more than one patient, the most frequently reported for BC development has been aberrations in *WWOX*. This tumour suppressor gene has been shown to be critical for normal breast development[34] with mutations in exons 4 to 9 frequently observed in BC tumours[35]. High expression of *WWOX* has been shown to be beneficial in association with tamoxifen treatment[52]. We further evaluated two unrelated fBC patients, one harbouring a CNV gain and the other a CNV loss. In both cases, the genomic rearrangements are predicted to reduce *WWOX* expression and thereby contribute to disease risk. Our results suggest that inherited deficiencies in *WWOX* are associated with disease but we could not demonstrate that these alterations were transmitted across generations due to ethical considerations. Notwithstanding, the frequency at which we have observed variants occurring in this gene (>1.55%) suggests that they may account for a significant proportion of *BRCA1*/*BRCA2* mutation negative fBC patients. Functional studies are required to determine the precise effect of these variants in the alteration of *WWOX* expression and BC development.

The identification of CNVs in close proximity to BC susceptibility genes and loci that either contributes to disease development directly or via more cryptic means expands our understanding of their contribution to disease risk in fBC. Our study identified CNVs residing in three genes *RPA3*, *NBN*, *MRE11A* and *CYP19A1* which supports their involvement in BC[6,28,29,53-56]. Given the predicted disruption of *RPA3*, *NBN*, *MRE11A* and *CYP19A1* it is likely that these variants are associated with disease.

Within our fBC cases we identified several genes within or in close proximity to rare CNVs which have previously been associated with BC: the putative oncogene *MLLT11* (aka *AF1Q*) has been reported to be over expressed in a BC cell line affecting invasive and metastatic potential[57,58]; while *PTK2B* has been shown to be the most frequently lost kinase in sporadic BC tumours and is suggested to contribute to the disease phenotype[59]. Of the rare CNVs associated with malignancy, the gene most frequently associated with BC development is the tumour suppressor *FHIT*. *FHIT* has been reported multiple times to be genetically and epigenetically modified in breast tumours[36-41]; its expression has been reported to be protective against *HER2*-driven breast tumour development[42]; whereas reduced expression is associated with poor prognosis[43]. A germline intronic deletion in *FHIT* has also been identified in a

pancreatic cancer study [60]. Given that we have found a constitutional CNV in *FHIT* we suggest that variants in this gene could also account for a fraction of fBC patients. As we were unable to obtain other family members it remains to be seen if these genomic re-arrangements confer significant disease risk in a family setting rather than being associated with disease progression.

A recent report using 68 patient and 100 controls suggested that rare CNVs may contribute to disease in a small proportion of fBC patients[24]. In contrast to our findings this study reported significantly lower percentages of rare CNVs in fBC patients (4%) compared to the level observed in the current study (30.65%)[24]. The discrepancies in these findings are most likely to be related to differences in sample populations, the type of array used (variation in array coverage and density), as well as the algorithm used by the analysis software[44-46]. These findings reinforce the need to obtain larger cohorts of patients and controls to better understand the contribution of CNVs to breast cancer development.

Conclusions

This study has revealed that there are a number of CNVs which may contribute to the development of fBC. Several previously reported BC susceptibility genes that include *RPA3*, *NBN*, *MRE11A* and *CYP19A1* were found to be influenced by the presence of a CNV. It was also revealed by this investigation that three unrelated fBC patients harboured CNVs in *WWOX* and *FHIT*. We propose that variants in these genes may account for disease in a significant proportion of fBC patients. Overall the results of this study provide further grounds for further investigation into the presence of CNVs in larger series of fBC patients who do not harbour changes in known breast cancer susceptibility genes.

List of Abbreviations

BC	Breast Cancer
bp	base pair
CGC	Cancer Gene Census
ChAS	Chromosome Analysis Suite (Affymetrix)
CN	Copy Number
CNV	Copy Number Variants
Cyto2.7M	Cytogenetic Whole Genome 2.7M array
DGV	Database of Genomic Variants
DSB	Double Strand Breat
DSBR	DSB Repair
fBC	familial Breast Cancer
HCS	Hunter Community Study
Kb	Kilobase
mapd	median of absolute pair-wise difference
Mb	Megabase
MLPA	Multiplex Ligation-dependant Probe Amplification
MMR	Mismatch Repair
NCG	Network of Cancer Genes
NTC	No template control
QC	quality control
SNP	single nucleotide polymorphism
WavinessSd	waviness standard deviation

Competing Interests

The authors declare that they have no competing interest.

Acknowledgements

This work has been supported by the following funding bodies and institutions: Australian Rotary Health/Rotary District 9650, the Commonwealth Scientific and Industrial Research Organization (CSIRO), the University of Newcastle and the Hunter Medical Research Institute. Samples were provided by the Hunter Area Pathology Service and the Hunter Community Study.

References

1. Jemal, A. *et al.* Global cancer statistics. *CA Cancer J Clin* **61**, 69-90 (2011).

2. Lalloo, F. & Evans, D.G. Familial breast cancer. *Clin Genet* **82**, 105-14 (2012).

3. Peto, J. & Mack, T.M. High constant incidence in twins and other relatives of women with breast cancer. *Nat Genet* **26**, 411-4 (2000).

4. Gracia-Aznarez, F.J. *et al.* Whole Exome Sequencing Suggests Much of Non-BRCA1/BRCA2 Familial Breast Cancer Is Due to Moderate and Low Penetrance Susceptibility Alleles. *PLoS One* **8**, e55681 (2013).

5. Wong, M.W. *et al.* BRIP1, PALB2, and RAD51C mutation analysis reveals their relative importance as genetic susceptibility factors for breast cancer. *Breast Cancer Res Treat* **127**, 853-9 (2011).

6. Zhang, B., Beeghly-Fadiel, A., Long, J. & Zheng, W. Genetic variants associated with breast-cancer risk: comprehensive research synopsis, meta-analysis, and epidemiological evidence. *Lancet Oncol* **12**, 477-88 (2011).

7. Michailidou, K. *et al.* Large-scale genotyping identifies 41 new loci associated with breast cancer risk. *Nat Genet* **45**, 353-61 (2013).

8. Palles, C. *et al.* Germline mutations affecting the proofreading domains of POLE and POLD1 predispose to colorectal adenomas and carcinomas. *Nat Genet* **45**, 136-44 (2013).

9. Lynch, H.T. & de la Chapelle, A. Genetic susceptibility to non-polyposis colorectal cancer. *J Med Genet* **36**, 801-18 (1999).

10. Kuiper, R.P. *et al.* Recurrence and variability of germline EPCAM deletions in Lynch syndrome. *Hum Mutat* **32**, 407-14 (2011).

11. Ligtenberg, M.J. *et al.* 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 (2009).

12. Chan, T.L. *et al.* A novel germline 1.8-kb deletion of hMLH1 mimicking alternative splicing: a founder mutation in the Chinese population. *Oncogene* **20**, 2976-81 (2001).

13. Nystrom-Lahti, M. *et al.* Founding mutations and Alu-mediated recombination in hereditary colon cancer. *Nat Med* **1**, 1203-6 (1995).

14. Stella, A. *et al.* Germline novel MSH2 deletions and a founder MSH2 deletion associated with anticipation effects in HNPCC. *Clin Genet* **71**, 130-9 (2007).

15. Plaschke, J., Ruschoff, J. & Schackert, H.K. Genomic rearrangements of hMSH6 contribute to the genetic predisposition in suspected hereditary non-polyposis colorectal cancer syndrome. *J Med Genet* **40**, 597-600 (2003).

16. Delnatte, C. *et al.* Contiguous gene deletion within chromosome arm 10q is associated with juvenile polyposis of infancy, reflecting cooperation between the BMPR1A and PTEN tumor-suppressor genes. *Am J Hum Genet* **78**, 1066-74 (2006).

17. van Hattem, W.A. *et al.* Large genomic deletions of SMAD4, BMPR1A and PTEN in juvenile polyposis. *Gut* **57**, 623-7 (2008).

18. Alonso-Espinaco, V. *et al.* Novel MLH1 duplication identified in Colombian families with Lynch syndrome. *Genet Med* **13**, 155-60 (2011).

19. Morak, M. *et al.* Biallelic MLH1 SNP cDNA expression or constitutional promoter methylation can hide genomic rearrangements causing Lynch syndrome. *J Med Genet* **48**, 513-519 (2011).

20. Clendenning, M. *et al.* Mutation deep within an intron of MSH2 causes Lynch syndrome. *Fam Cancer* **10**, 297-301 (2011).

21. Charames, G.S. *et al.* A large novel deletion in the APC promoter region causes gene silencing and leads to classical familial adenomatous polyposis in a Manitoba Mennonite kindred. *Hum Genet* **124**, 535-41 (2008).

22. Rohlin, A. *et al.* Inactivation of promoter 1B of APC causes partial gene silencing: evidence for a significant role of the promoter in regulation and causative of familial adenomatous polyposis. *Oncogene* (2011).

23. Giarola, M. *et al.* Screening for mutations of the APC gene in 66 Italian familial adenomatous polyposis patients: evidence for phenotypic differences in cases with and without identified mutation. *Hum Mutat* **13**, 116-23 (1999).

24. Krepischi, A.C. *et al.* Germline DNA copy number variation in familial and early-onset breast cancer. *Breast Cancer Res* **14**, R24 (2012).

25. Suehiro, Y. *et al.* Germline copy number variations associated with breast cancer susceptibility in a Japanese population. *Tumour Biol* **34**, 947-52 (2013).

26. McEvoy, M. *et al.* Cohort profile: The Hunter Community Study. *Int J Epidemiol* **39**, 1452-63 (2010).

27. Miller, S.A., Dykes, D.D. & Polesky, H.F. A simple salting out procedure for extracting DNA from human nucleated cells. *Nucleic Acids Res* **16**, 1215 (1988).

28. Murata, H., Khattar, N.H., Gu, L. & Li, G.M. Roles of mismatch repair proteins hMSH2 and hMLH1 in the development of sporadic breast cancer. *Cancer Lett* **223**, 143-50 (2005).

29. Vodusek, A.L., Novakovic, S., Stegel, V. & Jereb, B. Genotyping of BRCA1, BRCA2, p53, CDKN2A, MLH1 and MSH2 genes in a male patient with secondary breast cancer. *Radiol Oncol* **45**, 296-9 (2011).

30. Cancer Gene Census. (Cancer Genome Project Wellcome Trust Sanger Institute, 2012).

31. D'Antonio, M., Pendino, V., Sinha, S. & Ciccarelli, F.D. Network of Cancer Genes (NCG 3.0): integration and analysis of genetic and network properties of cancer genes. *Nucleic Acids Res* **40**, D978-83 (2012).

32. QuickCalcs - T test. (GraphPad Software Inc., GraphPad Software Inc., 2013).

33. Chang, N.S. *et al.* Hyaluronidase induction of a WW domain-containing oxidoreductase that enhances tumor necrosis factor cytotoxicity. *J Biol Chem* **276**, 3361-70 (2001).

34. Abdeen, S.K. *et al.* Wwox inactivation enhances mammary tumorigenesis. *Oncogene* **30**, 3900-6 (2011).

35. Ekizoglu, S., Muslumanoglu, M., Dalay, N. & Buyru, N. Genetic alterations of the WWOX gene in breast cancer. *Med Oncol* **29**, 1529-35 (2012).

36. Campiglio, M. *et al.* FHIT loss of function in human primary breast cancer correlates with advanced stage of the disease. *Cancer Res* **59**, 3866-9 (1999).

37. Cecener, G. *et al.* Importance of novel sequence alterations in the FHIT gene on formation of breast cancer. *Tumori* **93**, 597-603 (2007).

38. Iliopoulos, D. *et al.* Roles of FHIT and WWOX fragile genes in cancer. *Cancer Lett* **232**, 27-36 (2006).

39. Ismail, H.M., Medhat, A.M., Karim, A.M. & Zakhary, N.I. Multiple Patterns of FHIT Gene Homozygous Deletion in Egyptian Breast Cancer Patients. *Int J Breast Cancer* **2011**, 325947 (2011).

40. Ismail, H.M., Medhat, A.M., Karim, A.M. & Zakhary, N.I. FHIT gene and flanking region on chromosome 3p are subjected to extensive allelic loss in Egyptian breast cancer patients. *Mol Carcinog* **50**, 625-34 (2011).

41. Negrini, M. *et al.* The FHIT gene at 3p14.2 is abnormal in breast carcinomas. *Cancer Res* **56**, 3173-9 (1996).

42. Bianchi, F., Tagliabue, E., Menard, S. & Campiglio, M. Fhit expression protects against HER2-driven breast tumor development: unraveling the molecular interconnections. *Cell Cycle* **6**, 643-6 (2007).

43. Arun, B. *et al.* Loss of FHIT expression in breast cancer is correlated with poor prognostic markers. *Cancer Epidemiol Biomarkers Prev* **14**, 1681-5 (2005).

44. Dellinger, A.E. *et al.* Comparative analyses of seven algorithms for copy number variant identification from single nucleotide polymorphism arrays. *Nucleic Acids Res* **38**, e105 (2010).

45. Tsuang, D.W. *et al.* The effect of algorithms on copy number variant detection. *PLoS One* **5**, e14456 (2010).

46. Zhang, D. *et al.* Accuracy of CNV Detection from GWAS Data. *PLoS One* **6**, e14511 (2011).

47. Yoshihara, K. *et al.* Germline copy number variations in BRCA1-associated ovarian cancer patients. *Genes Chromosomes Cancer* **50**, 167-77 (2011).

48. Zhang, B., Kirov, S. & Snoddy, J. WebGestalt: an integrated system for exploring gene sets in various biological contexts. *Nucleic Acids Res* **33**, W741-8 (2005).

49. Ong, D.C. *et al.* LARG at chromosome 11q23 has functional characteristics of a tumor suppressor in human breast and colorectal cancer. *Oncogene* **28**, 4189-200 (2009).

50. Sathyanarayana, U.G. *et al.* Aberrant promoter methylation and silencing of laminin-5-encoding genes in breast carcinoma. *Clin Cancer Res* **9**, 6389-94 (2003).

51. Bajrami, I. *et al.* Synthetic lethality of PARP and NAMPT inhibition in triple-negative breast cancer cells. *EMBO Mol Med* **4**, 1087-96 (2012).

52. Gothlin Eremo, A. *et al.* Wwox expression may predict benefit from adjuvant tamoxifen in randomized breast cancer patients. *Oncol Rep* **29**, 1467-74 (2013).

53. Bartkova, J. *et al.* Aberrations of the MRE11-RAD50-NBS1 DNA damage sensor complex in human breast cancer: MRE11 as a candidate familial cancer-predisposing gene. *Mol Oncol* **2**, 296-316 (2008).

54. Heikkinen, K., Karppinen, S.M., Soini, Y., Makinen, M. & Winqvist, R. Mutation screening of Mre11 complex genes: indication of RAD50 involvement in breast and ovarian cancer susceptibility. *J Med Genet* **40**, e131 (2003).

55. Hsu, H.M. *et al.* Breast cancer risk is associated with the genes encoding the DNA double-strand break repair Mre11/Rad50/Nbs1 complex. *Cancer Epidemiol Biomarkers Prev* **16**, 2024-32 (2007).

56. Yuan, S.S. *et al.* Role of MRE11 in cell proliferation, tumor invasion, and DNA repair in breast cancer. *J Natl Cancer Inst* **104**, 1485-502 (2012).

57. Chang, X.Z. *et al.* Identification of the functional role of AF1Q in the progression of breast cancer. *Breast Cancer Res Treat* **111**, 65-78 (2008).

58. Li, D.Q. *et al.* Gene expression profile analysis of an isogenic tumour metastasis model reveals a functional role for oncogene AF1Q in breast cancer metastasis. *Eur J Cancer* **42**, 3274-86 (2006).

59. Naylor, T.L. *et al.* High resolution genomic analysis of sporadic breast cancer using array-based comparative genomic hybridization. *Breast Cancer Res* **7**, R1186-98 (2005).

60. Lucito, R. *et al.* Copy-number variants in patients with a strong family history of pancreatic cancer. *Cancer Biol Ther* **6**, 1592-9 (2007).

CHAPTER 6: *IN-SILICO* ANALYSIS OF GENES DISRUPTED BY A CNV IN HEREDITARY BREAST CANCER

Introduction

Copy number variants (CNVs) represent a form of genomic variation associated with the duplication (gain) or deletion (loss) of genetic material. In the event where one copy of the gene is lost, only one functional copy remains. If the remaining copy is inactivated, malignancy may arise[356]. Alternatively, genomic duplications may result in the overexpression of a gene and increase the growth advantage of that cell[212]. Therefore investigation of regions of genomic loss or gain may provide new insight into disease development and/or progression.

Several reports have recently provided evidence of CNVs as likely contributors to cancer, including hereditary breast cancer[285,290,357,358]. Furthermore, advances in bioinformatic analysis tools has provided the means to undertake more in-depth interrogation of large genetic datasets with the aim of deriving meaningful relationships which underpin disease development. This includes pathway enrichment analysis and microRNA (miR) annotation of genes lists derived from investigations such as CNV analysis[359-361]. The benefits of such approaches have been reported in a squamous lung cancer study which revealed cell cycle related genes and associated miRs (*miR-142-5p* and *miR-9*) that appear to contribute to the pathogenesis of this disease and are suggested to serve as good biomarkers[362].

This part of the study aims to undertake pathway analysis and miR annotation of gene lists derived from the CNV analysis of *BRCA1*/*BRCA2* mutation negative hereditary breast cancer patients. This analysis aims to provide further insight into biologically meaningful relationships that could underpin the pathogenesis of disease in this patient cohort.

Publication

Amy L. Masson, Bente A. Talseth-Palmer and Rodney J. Scott (2015) Interrogation of genes disrupted by copy number variants in familial breast cancer using a bioinformatics approach, *International Journal of Cancer Research and Diagnosis*.

Co-author statement

I attest that Research Higher Degree candidate Amy Louise Masson contributed to the above manuscript including involvement in the conception and design of the study; conducting the laboratory work, data analysis and interpretation; and preparation of the manuscript.

Co-author	Signature	Date
Bente A. Talseth-Palmer		
Rodney J. Scott		

Amy Louise Masson Date: 01/09/2015

Professor Robert Callister Date: 01/09/2015

Assistant Dean Research Training

Interrogation of genes disrupted by copy number variants in familial breast cancer using a bioinformatics approach

Amy L. Masson[1,2], Bente A. Talseth-Palmer[1,2] and Rodney J. Scott[1,2,3,*]

[1] Information Based Medicine Program, Hunter Medical Research Institute, University of Newcastle, Newcastle, New South Wales, 2305 Australia; E-mails: Bente.Talseth-Palmer@newcastle.edu.au (B.T-P.); Rodney.Scott@newcastle.edu.au (R.J.S.);

[2] Division of Molecular Medicine, Hunter Area Pathology Service, John Hunter Hospital, Newcastle, New South Wales, 2305 Australia

[3] School of Biomedical Sciences and Pharmacy, Faculty of Health, University of Newcastle, New South Wales, 2308 Australia

*Author to whom correspondence should be addressed; E-mail: amy.l.masson@uon.edu.au (A.L.M.); Tel.: +61 (2) 4042 0313; Fax: +61 (2) 4042 0030

Abstract

Background

Breast cancer is the most common malignancy in women worldwide with up to and 30% arising in a familial setting from which ~ 6% harbour mutations in *BRCA1*/*BRCA2*. Up to 80% of familial breast cancer patients remain without a genetic diagnosis suggesting that more cryptic changes within the genome may account for their disease. Copy number variants (CNVs), a form of structural genetic variation may be associated with disease development in a proportion of these patients.

Methods

WebGestalt pathway analysis and TAM miR annotation software were used to interrogate CNVs previously identified in 129 *BRCA1*/*BRCA2* mutation negative familial breast cancer patients. We examined 134 genes that were unique to patients compared to a control population; and second, a subset from the 134 gene set comprising 77 genes, termed “rare genes” due to their absence from the Database of Genomic Variants (DGV).

Results

Pathway analysis of the 134 genes revealed 21 KEGG-pathways and 14 cytoband regions enriched in the patient cohort while the analysis of the 77 rare genes revealed 20 KEGG-pathways, 7 cytoband regions and a single 3’UTR region to be enriched. MiR annotation of 2 miRs (miR-18a and miR-18b) suggested to target the enriched 3’UTR region, revealed an over-representation of miRs in the miR-17 family and miR-106a cluster (*p=0.0148* and *p=0.0111* respectively). A significant over-representation of miRs among the Human miR associated Disease Database (HMDD) categories for glioma (*p=0.0462*), lymphoma (*p=0.0296*) and toxoplasmosis (*p=0.0148*) was also revealed.

Conclusions

Bioinformatic analysis of genomic data is a powerful method of delineating biological relationships in large genomic datasets. This study has identified several candidate genes, loci and miRs associated with CNVs in 129 familial breast cancer *BRCA1*/*BRCA2* mutation negative patients which may contribute to disease risk. Further investigation is required to determine their precise involvement in familial breast cancer.

Key Words

familial breast cancer, copy number variation, bioinformatics, pathway analysis, microRNAs, high-penetrance genes, disease predisposition

Introduction

Breast cancer is the most commonly diagnosed malignancy in women and account for ~23% of all new cancer cases and 14% of all female cancer deaths[1]. Somewhere between 27% and 30% of breast cancers arise in a familial setting and are associated with an earlier age of disease presentation[2,3]. Somewhere between 5% and 7% of familial breast cancers are attributed to mutations in a single high penetrance gene such as *BRCA1*, *BRCA2*, *PALB2* and *TP53* while the remaining 20% to 22% of disease has been associated with one of several moderate penetrance genes that include *CHEK2*, *ATM*, *CASP8*, *CTLA4*, *NBN*, *CYP19A1*, *TERT* and *XRCC3*[2,4-7]. Large-scale genome-wide studies have also revealed numerous variants and genomic loci also contributing to breast cancer risk[8,9].

Currently *BRCA1* and *BRCA2* represent the most frequently tested breast cancer susceptibility genes, however they account for less than 20% of all familial breast cancer patients[7,10,11]. For many patients coming from breast cancer families no genetic diagnosis can be found. This suggests that either other genes are involved in disease risk or other genomic events may result in the inactivation of breast cancer susceptibility genes.

Copy number variants (CNVs) represent a form of genomic variation associated with the duplication (gain) or deletion (loss) of genetic material. In the event where one copy of the gene is lost, only one functional copy remains. If the remaining copy is inactivated malignancy may arise (for a review of LOH in cancer see[12]). Alternatively, genomic duplication may result in the overexpression of the gene and may lead to growth advantage[13]. Therefore investigation of regions of genomic loss or gain may provide insights into disease development and/or progression.

CNVs are associated with disease when the gain or loss occurs 'directly' (disruption of functional coding sequence) or 'indirectly' (via disruption to non-coding sequences that include promoter or intronic regions) with respect to a gene or via a third more cryptic association involving alteration of the distribution of epigenetic marks resulting in gene expression changes[14-21].

Several reports have recently provided evidence of CNVs as likely contributors to familial breast cancer[22,23]. This includes results from our own investigations identifying genomic rearrangements in *WWOX* and *FHIT*[24]. Other studies examining *BRCA1*/*BRCA2* mutation negative patients have reported 26 rare CNV variants that appear to contribute to disease[25] and the enrichment of rare CNVs disrupting pathways

responsible for maintaining genomic integrity[26], including the DNA double-strand break repair pathway known to be associated with breast cancer risk[27-29]. The evidence suggests that CNV screening could be a significant aid in determining disease risk in a small proportion of familial breast cancers[30].

Current bioinformatics analysis techniques provide the means to undertake more in-depth interrogation of large genetic datasets with the aim of deriving meaningful relationships to disease. This includes pathway enrichment analysis and microRNA (miR) retrieval through annotation of genes lists[31-33]. Of particular interest, miRs or small (20-22 nucleotides) non-coding RNAs (sncRNAs) are responsible for the regulation of gene expression through the targeting of mRNA products for cleavage/translational repression and since their identification have been reported to contribute to disease[34]. The benefits of such approaches have been observed in a squamous lung cancer (SLC) study which revealed cell cycle related genes and associated miRs *miR-142-5p* and *miR-9*[35] that appear to contribute to the pathogenesis of SLC suggesting they may serve as SLC biomarkers. As such, the identification of miRs may serve to benefit the development of diagnostic and prognostic biomarkers for breast cancer[36]. Bioinformatics tools such as the web based WebGestalt and TAM may therefore provide new insights into biologically meaningful relationships that underpin familial breast cancer.

Here we have used data obtained in an earlier familial breast cancer case-control study[24] to undertake pathway analysis and miR annotation as a means to delineate meaningful relationships that could provide further insight into the pathogenesis of breast cancer in this patient set. Specifically we have explored 134 genes that were revealed to be unique in patients compared to healthy controls; and interrogated a subset of 77 of these genes that were considered to be rare as judged by the Database of Genomic Variants (DGV).

Materials and Methods

The study was approved by the Hunter New England Human Research Ethics Committee and the University of Newcastle Human Research Ethics Committee. Data used in the current study has previously been characterized[24].

Patients

For this study 129 familial breast cancer patients provided samples. All familial breast cancer patients were clinically diagnosed with early-onset familial breast cancer, were referred for routine clinical diagnostic testing involving screening for mutations in *BRCA1* and/or *BRCA2*. Mutation screening was performed using Sanger Sequencing and Multiplex ligation-dependant probe amplification (MLPA). No mutations were identified in any of the patients used for the current study and are thus considered to be *BRCA1/BRCA2* mutation negative. The average age of disease diagnosis was 40.7 years of age.

Controls

A sample size of 40 controls from the Hunter Community Study (HCS)[37] were used in the current study. These samples were from healthy individuals aged >55 years who were cancer free at the time of sample collection.

Genomic Array Preparation and Analysis

The DNA from the 169 patients and controls was processed on the Affymetrix Cytogenetic Whole Genome 2.7M (Cyto 2.7M) array according to manufacturer's protocols as described previously[24].

Statistical, pathway and annotation analysis

Gene enrichment analysis was performed using WebGestalt analysis software (Version 2013)[33]. This software was used to assess gene lists derived from the refined CNV results obtained from ChAS according to Kyoto Encyclopedia of Genes and Genomes (KEGG) pathways, cytoband enrichment and miR targets. Analysis was performed using hypergeometric statistical method, Benjamini and Hochberg (BH) correction for multiple testing (both default settings) and a biological significance threshold of <0.05 with a minimum of two genes per category required to assess any enrichment. TAM (Tool for Annotations of miRs) (Version 2)[31] software was available for use to annotate miRs according to miR family, cluster, function, Human miR associated disease database (HMDD) and tissue specificity. Annotations were performed using the

following parameters: all miRs in the TAM database were used as a background; to identify meaningful categories we looked at miR over-representation in all categories and analysis was limited to at least one miR in a given category. Enrichment analysis for miRs categories was conducted using hypergeometric testing and *p* values were corrected according to Bonferroni correction for multiple testing.

Results

The CNV data used in the current study has been partially characterized[24]. Briefly, 414 CNVs were identified across the 169 participants in this study, of which 310 CNVs were detected in the 129 breast cancer patients and an additional 104 were detected in the 40 controls. The size of CNVs observed ranged from 9.65 Kb to 1335.06 Kb in length. The average number of CNVs, average genomic burden and average CNV size did not significantly differ between patients and controls (p=0.75, p=0.30 and p=0.07 respectfully).

In the control cohort, 66 of the 104 CNVs (63.46%) disrupted 96 genes. These 96 genes were compared against the current cancer genome census list (COSMIC database available: http://cancer.sanger.ac.uk/cancergenome/projects/cosmic/) to ascertain if any were reported to be associated with cancer. While none of the 96 genes are known to be associated with a cancer predisposition, three CNVs were encompassing *MLL3*, *PBX1* and *PLAG1*, respectively, have been linked to medulloblastoma, pre-B-cell ALL/myoepithelioma and salivary adenoma.

Genes associated with CNVs unique to patients

A total of 275 CNVs were found to be unique to the breast cancer population and over half of these disrupted a similar number of genes. A CNV in *SUPT3H* was also removed from further analysis having been identified in a control sample to be affected by a rearrangement and considered unlikely to be associated with disease risk. The remaining 149 genes which were disrupted by a CNV in 65 patients had not hitherto been assessed for their involvement in breast cancer. WebGestalt [33] pathway analysis software mapped 134 of the 149 genes and these were compared to all genes in the human genome to search for enriched KEGG-pathways, cytobands and miR targets (i.e. 3'UTR regions of genes; see supplementary table 1).

KEGG analysis revealed 21 significant pathways in which 38 genes uniquely identified in the patients were enriched (see table 1). These included: pathways-in-cancer (p=0.0011), endometrial cancer (p=0.0040), the tight junction pathway (p=0.0040), small cell lung cancer (p=0.0080), chemokine signalling pathway (p=0.0087), leukocyte transendothelial migration (p=0.0145), aldosterone-regulated sodium reabsorption (p=0.0175), insulin signalling pathway (p=0.0199), non-small cell lung cancer (p=0.0246), B-cell receptor signalling pathway (p=0.0343) and renal cell carcinoma (p=0.0343) pathway as well as the adherens junction pathway (p=0.0343). Of particular interest was the enrichment of 13 of the 21 pathways all containing *PIK3R5*

for which 3 patients harboured CNVs disrupting exonic and intronic gene sequences. The first patient harboured the largest of the CNVs, located 17:8,736,773-8,820,505 (84 Kb gain, detected by 68 probes; 91% confidence). The second patient harboured a CNV located 17:8,738,684-8,800,866 (62 Kb gain, detected by 52 probes; 90% confidence), while the third patient harboured the smallest CNV located 17:8,756,373-8,786,411 (30 Kb gain, detected by 26 probes; 91% confidence). All these CNVs overlapped the same 30 Kb region defined by the smallest of the three CNVs identified. Furthermore, 7 of the 21 enriched pathways similarly contained the same genes, *CTNNA2* and *CTNNA3*, which were disrupted by a CNV in one patient each. The CNV disrupting *CTNNA2* was located 2:80,233,408-80,307,154 (74 Kb loss, detected by 101 probes; 91% confidence) while the CNV disrupting *CTNNA3* was located 10:67,744,533-67,785,552 (41 Kb loss, detected by 48 probes; 93% confidence) both affecting only intronic sequences of the respective genes. Overall, the KEGG-pathway in which the most patients harboured CNVs disrupting the KEGG-descriptor 'pathways *in cancer*', which encompassed *PIK3R5*, *CTNNA2* and *CTNNA3* in addition to *EGLN3*, *TFG*, *IL8* and *LAMB3* which was identified in 7 patients (with 2 harbouring CNVs in more than one of the described genes). Two patients harboured CNVs disrupting exonic and intronic sequence of *LAMB3*, located 1:207,852,879-207,924,028 (71Kb gain, detected by 32 probes; 92% confidence) and 1:207,890,364-207,939,169 (49 Kb gain, detected by 34 probes; 90% confidence), respectively, and overlapped the same region 49 Kb in size. CNVs disrupting *IL8* also affected 2 patients, and were located 4:74,788,136-74,829,138 (41 Kb gain, detected by 47 probes; 93% confidence) and 4:74,788,969-74,846,104 (57 Kb, detected by 68 probes; 91% confidence). These CNVs overlapped a region of 40 Kb and disrupted both exonic and intronic gene sequence in both patients. One other patient harboured a whole gene duplication of *EGLN3* located 14:33,359,188-33,601,677 (242 Kb, detected by 160 probes; 92% confidence), while another harboured a CNV gain of exonic and intronic gene sequence in *TFG* located 3:101,822,318-101,928,257 (106 Kb, detected by 97 probes; 93% confidence). The identification of genes associated with cancer pathways disrupted by a CNV in our cohort suggests the possible involvement of these genes in the development of breast cancer.

Table 1 Summary of WebGestalt analysis showing enriched KEG-pathways, genes observed in the 143 gene set unique to the fBC patients and the significance value (*p*).

Description	Genes	*p*
Systemic lupus erythe-matosus	*HIST1H2AD* *HIST1H2BD* *HIST1H2AE* *HIST1H2BE* *HIST1H2BF* *HIST1H2BH* *HIST1H2BG* *HIST1H2BI* *HIST1H3D* *HIST1H3F* *HIST1H3E* *HIST1H3G* *HIST1H4D* *HIST1H4F* *HIST1H4E* *HIST1H4G*	0.0000
Pathways in cancer	*CTNNA2* *CTNNA3* *LAMB3* *PIK3R5* *EGLN3* *IL8* *TFG*	0.0011
Endometrial cancer	*CTNNA3* *CTNNA2* *PIK3R5*	0.0040
Glycerolipid metabolism	*DGKB* *GPAT2* *MBOAT1*	0.0040
Tight junction	*CTNNA3* *CTNNA2* *MYH8* *MYH13*	0.0040
Staphylococcus aureus infection	*C3AR1* *FPR1* *FPR2*	0.0040
Bacterial invasion of epithelial cells	*CTNNA3* *CTNNA2* *PIK3R5*	0.0060
Glycerophos-pholipid metabolism	*DGKB* *GPAT2* *MBOAT1*	0.0075
Small cell lung cancer	*FHIT* *LAMB3* *PIK3R5*	0.0080
Chemokine signalling pathway	*ADRBK2* *GNG2* *IL8* *PIK3R5*	0.0087
Amoebiasis	*IL8* *LAMB3* *PIK3R5*	0.0123
Leukocyte transendo-thelial migration	*CTNNA3* *CTNNA2* *PIK3R5*	0.0145
Aldosterone-regulated sodium	*PIK3R5* *HSD11B1*	0.0175

reabsorption			
Insulin signalling pathway	*PIK3R5* *PYGL*	*PHKB*	0.0199
Non-small cell lung cancer	*FHIT*	*PIK3R5*	0.0246
Renal cell carcinoma	*EGLN3*	*PIK3R5*	0.0343
Arrhyth-mogenic right ventricular cardio-myopathy	*CTNNA3*	*CTNNA2*	0.0343
B cell receptor signalling pathway	*CARD11*	*PIK3R5*	0.0343
Adherens junction	*CTNNA3*	*CTNNA2*	0.0343
Viral myocarditis	*MYH13*	*MYH8*	0.0343
Phosphat-idylinositol signalling system	*DGKB*	*PIK3R5*	0.0353

Enrichment analysis revealed 14 cytobands of significance (2p22, 3q12, 3p26, 5q12, 6p, 6p21, 6p22, 7q31, 10q23, 11q21, 12p, 17q11 and 21q; all with $p<0.0436$) containing 56 genes that represent candidate genes for familial breast cancer (see table 2). The cytoband 21q in which *C21orf91*, *DSCAM*, *NCAM2*, *C21orf7* and *KCNJ15* reside were disrupted by CNVs in 4 patients and represented the most common enriched cytoband to be affected by CNVs. The first patient harboured three CNVs at this cytoband, located *C21orf91*, 21:18,074,850-18,090,874 (16 Kb gain, detected by 24 probes; 91% confidence), *C21orf7*, 21:29,468,330-29,497,829 (30 Kb gain, detected by 28 probes; 91% confidence) and *KCNJ15*, 21:38,558,908-38,600,216 (41 Kb gain, detected by 58 probes; 90% confidence). A second patient harboured a CNV gain located *NCAM2*, 21:21,268,131-21,367,798 (100 Kb, detected by 114 probes; 94% confidence). The third and fourth patients both harboured CNV gains disrupting *DSCAM*, located 21:40,624,405-40,653,316 (29 Kb, detected by 37 probes; 94% confidence) and 21:40,631, 461-40655,668 (24 Kb, detected by 32 probes; 92% confidence), both overlapping the same 22 Kb region. All CNVs affected exonic and intronic gene sequences. Together these results suggest a possible role of the enriched genes at this cytoband in breast cancer.

Table 2 Summary of WebGestalt analysis showing enriched cytobands, genes observed in the 143 gene set unique to the fBC patients and the significance value (*p*).

Description	Genes		*p*
6p22	*HIST1H2AE* *HIST1H2BE* *HIST1H3E* *HIST1H3F* *HIST1H4E* *HIST1H4F* *HIST1H4G*	*HIST1H2BF* *HIST1H2BI* *HIST1H3D* *HIST1H3G* *HIST1H4D* *MBOAT1*	0.0000
6p	*HIST1H1D* *HIST1H1E* *HIST1H2BD* *HIST1H2BE* *HIST1H2BF* *HIST1H2BI* *HIST1H3D* *HIST1H3E* *HIST1H4E* *HIST1H4F*	*HIST1H2AD* *HIST1H2AE* *HIST1H2BG* *HIST1H2BH* *HIST1H3F* *HIST1H3G* *HIST1H4D* *HIST1H4G* *KIAA1586* *MBOAT1*	0.0000
2p22	*BIRC6* *GALM* *HNRPLL*	*LOC100271832* *TTC27*	0.0000
17q11	*CPD* *EVI2A*	*EVI2B* *NF1*	0.0124
7q31	*DOCK4* *IMMP2L*	*LRRN3*	0.0413
3p26	*CHL1*	*CNTN4*	0.0413
12p	*ARHGDIB* *C3AR1* *ERP27*	*FOXJ2* *NECAP1*	0.0436
3q12	*GPR128*	*TFG*	0.0436
11q21	*FUT4*	*PIWIL4*	0.0436
5q12	*IPO11*	*LRRC70*	0.0436
21q	*C21orf7*	*KCNJ15*	0.0436

	C21orf91 *DSCAM*	*NCAM2*	
10q23	*FAM22A* *FAM35A*	*LOC728190*	0.0436
6p21	*HIST1H1E* *HIST1H1D* *HIST1H2BG*	*HIST1H2BD* *HIST1H2BH* *HIST1H2AD*	0.0436
13q12	*ALOX5AP* *MPHOSPH8*	*PSPC1*	0.0436

Screening of the 3'UTR region of target genes of the miRs, failed to identify any significant regions that were patient specific.

Rare genes associated with familial breast cancer

Of the 180 CNVs associated with genes, a total of 70, found in 28 of the patients, had not been reported in the DGV. These 70 CNVs were associated with 77 genes which were investigated further using WebGestalt pathway analysis software[33].

KEGG analysis revealed a total of 20 significant pathways in which 18 genes were enriched (see table 3). Among those pathways identified were pathways involved in small cell lung cancer (p=0.0060), leukocyte transendothelial migration (p=0.0060), the spliceosome (p=0.0067), aldosterone-regulated sodium re-absorption (p=0.0100), non-small cell lung cancer (p=0.0137), B-cell signalling pathway (p=0.0203), hepatitis C (p=0.0290), pathways in cancer (p=0.0290), the insulin signalling pathway (p=0.0060), the T-cell receptor signalling pathway (p=0.0060), the tight junction pathway (p=0.0290) and the toll-like receptor signalling pathway (p=0.0060) were identified to be among those enriched. In particular 16 of the 21 pathways were enriched by the presence of the same gene, *PIK3R5*, which was disrupted by a CNV in 3 patients (see above) while a second unrelated pathway '*vascular smooth muscle contraction*' containing *ARHGEF12* was disrupted by a CNV in 4 other patients. The first patient contained a CNV located 11:119,679,290-119,726,037 (47 Kb gain, detected by 52 probes; 91% confidence) while the other three patients contained smaller CNVs located 11:119,691,589-119,728,609 (37 Kb gain, detected by 43 probes; 91% confidence), 11:119,695,754-119,724,126 (28 Kb gain, detected by 36 probes; 92% confidence) and 11:119,697,081-119,723,342 (26 Kb gain, detected by 33 probes; 91% confidence). All four CNVs affected a similar region of 26 Kb (as defined by the smallest of the detected CNVs) and disrupted both exonic and intronic gene sequence. The high frequency of patients affected by a CNV disrupting the same gene suggests the possible involvement of *PIK3R5* and *ARHGEF12* in disease development in the affected 7 patients.

Table 3 Summary of WebGestalt analysis showing enriched KEGG-pathways, genes observed in the 77 gene set of rare genes unique to the fBC patients and the significance value (*p*).

Description	Genes		*p*
Small cell lung cancer	*FHIT*	*LAMB3*	0.006
	PIK3R5		
Chemokine signalling pathway	*GNG2*	*IL8*	0.006
	PTK2B	*PIK3R5*	
Leukocyte transendothelial migration	*PTK2B*	*PIK3R5*	0.0066
	RHOH		
Amoebiasis	*IL8*	*LAMB3*	0.0066
	PIK3R5		
Spliceosome	*PCBP1*	*SLU7*	0.0067
	SNRNP27		
Aldosterone-regulated sodium reabsorption	*HSD11B1*	*PIK3R5*	0.01
Non-small cell lung cancer	*FHIT*	*PIK3R5*	0.0137
Viral myocarditis	*MYH8*	*MYH13*	0.0201
B cell receptor signalling pathway	*CARD11*	*PIK3R5*	0.0203
Hepatitis C	*IL8*	*PIK3R5*	0.029
Toxoplasmosis	*LAMB3*	*PIK3R5*	0.029
Pathways in cancer	*IL8*	*LAMB3*	0.029
	PIK3R5		
Insulin signalling pathway	*PIK3R5*	*PYGL*	0.029
Natural killer cell mediated cytotoxicity	*PIK3R5*	*PTK2B*	0.029
Chagas disease (American trypanosomiasis)	*IL8*	*PIK3R5*	0.029
T cell receptor signalling pathway	*CARD11*	*PIK3R5*	0.029
Tight junction	*MYH8*	*MYH13*	0.029
Vascular smooth muscle contraction	*ARHGEF12*	*PRKG1*	0.029
Neurotrophin signalling pathway	*ARHGDIB*	*PIK3R5*	0.029
Toll-like receptor signalling pathway	*IL8*	*PIK3R5*	0.029

Furthermore, cytogenetic band enrichment analysis identified 7 cytobands (2p13, 11q21, 11q23, 2p16, 12p12, 1q32 and 5q33; all with $p<0.0412$) which were associated with 17 of the rare genes (see table 4). The cytoband 11q23 in which *ARHGEF12*, *TMEM136* and *POU2F3* reside were disrupted by CNVs in 4 patients (individual CNVs described earlier) in addition to the cytoband 1q32 containing *G0S2* and *LAMB3* which were disrupted by CNVs in 2 other patients (also described earlier). These results further suggest a possible role of *G0S2* and *LAMB3* in breast cancer.

Table 4 Summary of WebGestalt analysis showing enriched cytobands, genes observed in the 77 gene set of rare genes unique to the fBC patients and the significance value (*p*).

Description	Genes			*p*
2p13	*ASPRV1*	*LOC100133985*	*SNRNP27*	0.0279
11q21	*FUT4*	*PIWIL4*		0.0372
11q23	*ARHGEF12*	*POU2F3*	*TMEM136*	0.0372
2p16	*ACYP2*	*TSPYL6*		0.0412
12p12	*ARHGDIB*	*ERP27*		0.0412
1q32	*CAMK1G*	*G0S2*	*LAMB3*	0.0412
5q33	*SLU7*	*C5orf54*		0.0412

Enrichment analysis for miR targets identified one significant region within the 3'UTR of 4 of the rare genes (hsa_GCACCTT in *PSD3*, *ARL15*, *NEDD9* and *SMAP2*; p=0.0018; see table 5). A total of 3 patients contained CNVs associated with the 4 genes containing the enriched region. The first patient contained CNVs affecting two of the genes, located in the intronic and exonic sequence of *NEDD9*, 6:11,428,833-11,512,036 (83 Kb gain, detected by 48 probes; 91% confidence) and the intronic and exonic sequence of *SMAP2*, 1:40,622,494-40,649,739 (27 Kb gain, detected by 38 probes; 91% confidence). A second patient harboured a CNV loss within the exonic and intronic sequence of *PSD3* located 8:18,752,814-18,769,132 (16 Kb, detected by 26 probes; 91% confidence) while a third patient harboured an intronic CNV gain in *ARL15* located 5:53,231,517-53,299,530 (68 Kb, detected by 60 probes; 94% confidence). A total of two miRs, miR-18a and miR-18b, were proposed to target the enriched region. TAM [31] was used to identify over-representation of miR categories as a result of the two miRs. We identified a total of 16 miR categories: 1 family, 1 cluster, 4 functional categories, 10 HMDD and no tissue specificity categories. Significance was attained for one miR family; miR-17 (miR-18b; p=0.0148); one cluster, miR-106a (miR-18b; *p*=0.0111); one functional group, immune system (*p*=0.0333) and three HMDDs, being glioma (*p*=0.0462), lymphoma (*p*=0.0296) and toxoplasmosis (*p*=0.0148).

Table 5 Summary of WebGestalt analysis showing enriched 3'UTR regions, proposed miR targets for the 3'UTR region, genes observed in the 77 gene set of rare genes unique to the fBC patients and the significance value (*p*).

Description	Genes	*p*
hsa_GCACCTT (mir-18A, miR-18B)	*ARL15* *NEDD9* *PSD3* *SMAP2*	0.0018

Discussion

We have undertaken *in-silico* analysis using CNV data derived from the Cyto2.7M array from 129 familial breast cancer *BRCA1*/*BRCA2* mutation negative patients to search for meaningful relationships among genes, genomic regions and miR controlling species which may provide new insights into their potential contribution to disease development.

The Cancer Genome Atlas (TCGA) identifies 17 of the 134 genes analysed in the current study as being disrupted by a CNVs in 773 breast tumours; including *ALOX5AP*, *ARHGDIB*, *B2M*, *CARD11*, *DMXL2*, *ERP27*, *FAM135B*, *FHIT*, *GPR39*, *LRP5L*, *MALAT5*, *MPHOSPH8*, *NEDD9*, *PHKB*, *PSPC1*, *SCYL1* and *TRIM69I*[38] reinforcing their likely involvement in disease development in these patients.

Enrichment analysis of KEGG-pathways, cytoband regions and miR targets was conducted for (1) genes identified unique to patients compared to regional controls; and (2) genes considered to be rare when compared to the Database of Genomic Variants (DGV). The purpose of these two analyses was to further determine the significance of rare genomic variants and their contribution to disease risk. The benefit of examining rare CNVs is that the regions encompassed by the variation are proposed to harbour genes or other regulatory elements that are likely to be significant with respect to disease risk[22,23].

Analysis of the 134 genes associated with CNVs unique to the patients identified 21 enriched KEGG-pathways. Several pathways most likely to be involved in the aetiology of breast cancer were driven by 38 genes representing potential genetic risk candidates. Enriched KEGG-pathways included endometrial cancer, renal cell carcinoma and the adherens junction pathway. These pathways have demonstrated links to cancer development and progression[39-42]. For example phosphorylation of the adherens junction protein Afadin has been reported to results in increased breast cancer cell migration and cancer progression[41].

Many of the enriched KEGG-pathways (a total of 11) which included aldosterone-regulated sodium reabsorption, amoebiasis, B-cell receptor signalling pathway, pathways in cancer, chemokine signalling, leukocyte transendothelial migration, insulin signalling pathway, small cell lung cancer, non-small cell lung cancer, the tight junction pathway and the viral myocarditis pathway also represented KEGG-pathways that were enriched among rare genes indicating that the rare genes are driving many of the observed associations. These pathways have also been reported to contribute to

malignancy[43-53], for example low levels of tight junction plaque molecules (*ZO-1* and *MUPP-1*) are associated with poor prognosis in breast cancer patients[48] .

Interrogation of rare genes specifically has further revealed enriched KEGG-pathways related to the spliceosome, hepatitis C, T-cell receptor signalling and toll-like receptor signalling which were not enriched in the analysis of all unique genes. These pathways have been reported to have associations with breast cancer[54-58], for example breast cancer patients who have hepatitis C infection demonstrate poorer clinical outcomes to breast cancer treatment[56]. A total of 17 rare genes were associated with the enriched KEGG-pathways and represent candidate high penetrance genes in breast cancer risk.

Interrogation of both 134 unique genes and 77 rare genes for enriched cytobands led to the identification of 20 cytoband regions. Specifically of the 14 cytoband regions arising from the analysis of the unique genes alone, at least five have been identified as risk loci for breast cancer in recent GWASs and meta-analyses (3p26, 6p, 6p22, 12p and 21q)[9,59-64]. Associated with the 14 cytobands were 56 genes representing additional candidate breast cancer susceptibility genes. Only one of the cytobands (11q21) was identified to be in common with a cytoband revealed by the interrogation of the rare genes alone, in which six additional cytobands (1q32, 2p13, 2p16, 5q33, 11q23 and 12p12) were also discovered. Interestingly three (2p16, 1q32 and 5q33) of these regions have also been reported as breast cancer susceptibility regions[9,59,60,62,65]. These cytobands were associated with 17 genes which can be considered candidates for breast cancer susceptibility loci.

Enrichment analysis for the targets of miRs led to the discovery of two miRs which are proposed to target the 3'UTR of four rare genes revealed by our CNV analysis. These four genes represent potential candidate genes for contribution to familial breast cancer. The annotation of the two miRs in TAM[31] yielded an over representation of miRs in miR-17 family and miR-106a category, both containing the miR miR-18b. A recent study has reported a significant association between high expression of miR-18a/b in basal-like breast cancer[66] and another study has shown that low expression of miR-18b was correlated with improved survival in HER2-negative breast cancers [67]. The immune system was over-represented in the functional group analysis and is well documented with respect to its association with breast cancer (see[68] for review). Furthermore, a recent study has also suggested macrophage and T-cell abundance in breast cancer to be prognostic indicators for recurrence-free and overall patient survival[45]. Lastly, among the HMDD analysis, glioma, toxoplasmosis and lymphoma were overrepresented. The overrepresentation of toxoplasmosis and lymphoma

categories represented an unexpected finding as they are rarely associated with breast cancer[69,70]. With respect to the current study, our results suggests that miR-18a and miR-18b shows potential relationships with disrupted genes in the familial breast cancer cohort and may therefore play a role in familial breast cancer.

A potential pitfall of this analysis remains the relatively poor annotation of disease associated pathways. This study has relied on the best available evidence however this still remains an area in need of further development before we fully understand the natural relationships between CNVs and all the pathways involved in cancer initiation and progression.

In summary, biological attributes of both rare and unique CNVs identified in familial breast cancer patients was investigated using pathway analysis resulting in the identification of several networks, miRs and cytoband regions both previously implicated in cancer development and some novel findings which are potentially involved in familial breast cancer risk. Overall the results of this study provide *in-silico* evidence for the involvement of CNVs in the aetiology of familial breast cancer.

List of abbreviations

BH	Benjamini and Hochberg
ChAS	Chromosome Analysis Suite (Affymetrix)
CN	Copy Number
CNV	Copy Number Variants
Cyto2.7M	Cytogenetic Whole Genome 2.7M array
DGV	Database of Genomic Variants
HCS	Hunter Community Study
HMDD	Human microRNA associated Diseases
KEGG	Kyoto Encyclopaedia of Genes and Genomes
mapd	median of absolute pair-wise difference
miR	microRNA
MLPA	Multiplex Ligation-dependant Probe Amplification
QC	quality control
SLC	squamous lung cancer
sncRNA	short non-coding RNA
SNP	single nucleotide polymorphism
TAM	Tool for the Annotation of MicroRNAs
UTR	Un-Transcribed Region
Waviness Sd	waviness standard deviation

Competing interests

The authors declare that they have no competing interest.

Authors contributions

ALM conducted the experiments, performed data analysis/interpretation and wrote the first draft of the manuscript.

BAT-P, GNH provided critical review of the manuscript.

RJS conceived the study and reviewed and approved the final version of the manuscript prior to submission.

Acknowledgements

This work has been supported by the following funding bodies and institutions: Australian Rotary Health/Rotary District 9650, the Commonwealth Scientific and Industrial Research Organization (CSIRO), the University of Newcastle and the Hunter Medical Research Institute. Samples were provided by the Hunter Area Pathology Service and the Hunter Community Study.

References

1. Jemal, A. *et al.* Global cancer statistics. *CA Cancer J Clin* **61**, 69-90 (2011).

2. Lalloo, F. & Evans, D.G. Familial breast cancer. *Clin Genet* **82**, 105-14 (2012).

3. Peto, J. & Mack, T.M. High constant incidence in twins and other relatives of women with breast cancer. *Nat Genet* **26**, 411-4 (2000).

4. Gracia-Aznarez, F.J. *et al.* Whole Exome Sequencing Suggests Much of Non-BRCA1/BRCA2 Familial Breast Cancer Is Due to Moderate and Low Penetrance Susceptibility Alleles. *PLoS One* **8**, e55681 (2013).

5. Wong, M.W. *et al.* BRIP1, PALB2, and RAD51C mutation analysis reveals their relative importance as genetic susceptibility factors for breast cancer. *Breast Cancer Res Treat* **127**, 853-9 (2011).

6. Mavaddat, N., Antoniou, A.C., Easton, D.F. & Garcia-Closas, M. Genetic susceptibility to breast cancer. *Mol Oncol* **4**, 174-91 (2010).

7. Stratton, M.R. & Rahman, N. The emerging landscape of breast cancer susceptibility. *Nat Genet* **40**, 17-22 (2008).

8. Zhang, B., Beeghly-Fadiel, A., Long, J. & Zheng, W. Genetic variants associated with breast-cancer risk: comprehensive research synopsis, meta-analysis, and epidemiological evidence. *Lancet Oncol* **12**, 477-88 (2011).

9. Michailidou, K. *et al.* Large-scale genotyping identifies 41 new loci associated with breast cancer risk. *Nat Genet* **45**, 353-61 (2013).

10. Nathanson, K.L., Wooster, R. & Weber, B.L. Breast cancer genetics: what we know and what we need. *Nat Med* **7**, 552-6 (2001).

11. Wooster, R. & Weber, B.L. Breast and ovarian cancer. *N Engl J Med* **348**, 2339-47 (2003).

12. Lasko, D., Cavenee, W. & Nordenskjold, M. Loss of constitutional heterozygosity in human cancer. *Annu Rev Genet* **25**, 281-314 (1991).

13. Harewood, L., Chaignat, E. & Reymond, A. Structural variation and its effect on expression. *Methods Mol Biol* **838**, 173-86 (2012).

14. Chan, T.L. *et al.* A novel germline 1.8-kb deletion of hMLH1 mimicking alternative splicing: a founder mutation in the Chinese population. *Oncogene* **20**, 2976-81 (2001).

15. Stella, A. *et al.* Germline novel MSH2 deletions and a founder MSH2 deletion associated with anticipation effects in HNPCC. *Clin Genet* **71**, 130-9 (2007).

16. van Hattem, W.A. *et al.* Large genomic deletions of SMAD4, BMPR1A and PTEN in juvenile polyposis. *Gut* **57**, 623-7 (2008).

17. Alonso-Espinaco, V. *et al.* Novel MLH1 duplication identified in Colombian families with Lynch syndrome. *Genet Med* **13**, 155-60 (2011).

18. Clendenning, M. *et al.* Mutation deep within an intron of MSH2 causes Lynch syndrome. *Fam Cancer* **10**, 297-301 (2011).

19. Charames, G.S. *et al.* A large novel deletion in the APC promoter region causes gene silencing and leads to classical familial adenomatous polyposis in a Manitoba Mennonite kindred. *Hum Genet* **124**, 535-41 (2008).

20. Rohlin, A. *et al.* Inactivation of promoter 1B of APC causes partial gene silencing: evidence for a significant role of the promoter in regulation and causative of familial adenomatous polyposis. *Oncogene* (2011).

21. Ligtenberg, M.J. *et al.* 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 (2009).

22. Conrad, D.F. *et al.* Origins and functional impact of copy number variation in the human genome. *Nature* **464**, 704-12 (2010).

23. Shlien, A. & Malkin, D. Copy number variations and cancer susceptibility. *Curr Opin Oncol* **22**, 55-63 (2010).

24. Masson, A.L. *et al.* Expanding the genetic basis of copy number variation in familial breast cancer. *Hered Cancer Clin Pract* **12**, 15 (2014).

25. Krepischi, A.C. *et al.* Germline DNA copy number variation in familial and early-onset breast cancer. *Breast Cancer Res* **14**, R24 (2012).

26. Pylkas, K. *et al.* Rare copy number variants observed in hereditary breast cancer cases disrupt genes in estrogen signaling and TP53 tumor suppression network. *PLoS Genet* **8**, e1002734 (2012).

27. Bau, D.T., Mau, Y.C., Ding, S.L., Wu, P.E. & Shen, C.Y. DNA double-strand break repair capacity and risk of breast cancer. *Carcinogenesis* **28**, 1726-30 (2007).

28. Sehl, M.E. *et al.* Associations between single nucleotide polymorphisms in double-stranded DNA repair pathway genes and familial breast cancer. *Clin Cancer Res* **15**, 2192-203 (2009).

29. Wu, H.C. *et al.* DNA double-strand break repair genotype and phenotype and breast cancer risk within sisters from the New York site of the Breast Cancer Family Registry (BCFR). *Cancer Causes Control* **24**, 2157-68 (2013).

30. Suehiro, Y. *et al.* Germline copy number variations associated with breast cancer susceptibility in a Japanese population. *Tumour Biol* **34**, 947-52 (2013).

31. Lu, M., Shi, B., Wang, J., Cao, Q. & Cui, Q. TAM: a method for enrichment and depletion analysis of a microRNA category in a list of microRNAs. *BMC Bioinformatics* **11**, 419 (2010).

32. Wang, J., Duncan, D., Shi, Z. & Zhang, B. WEB-based GEne SeT AnaLysis Toolkit (WebGestalt): update 2013. *Nucleic Acids Res* **41**, W77-83 (2013).

33. Zhang, B., Kirov, S. & Snoddy, J. WebGestalt: an integrated system for exploring gene sets in various biological contexts. *Nucleic Acids Res* **33**, W741-8 (2005).

34. Di Leva, G., Garofalo, M. & Croce, C.M. MicroRNAs in cancer. *Annu Rev Pathol* **9**, 287-314 (2014).

35. Su, Y.H. *et al.* MIR-142-5p and miR-9 may be involved in squamous lung cancer by regulating cell cycle related genes. *Eur Rev Med Pharmacol Sci* **17**, 3213-20 (2013).

36. Macha, M.A. *et al.* MicroRNAs (miRNA) as Biomarker(s) for Prognosis and Diagnosis of Gastrointestinal (GI) Cancers. *Curr Pharm Des* (2014).

37. McEvoy, M. *et al.* Cohort profile: The Hunter Community Study. *Int J Epidemiol* **39**, 1452-63 (2010).

38. Cancer Genome Atlas, N. Comprehensive molecular portraits of human breast tumours. *Nature* **490**, 61-70 (2012).

39. Fisher, B. *et al.* Endometrial cancer in tamoxifen-treated breast cancer patients: findings from the National Surgical Adjuvant Breast and Bowel Project (NSABP) B-14. *J Natl Cancer Inst* **86**, 527-37 (1994).

40. Segev, Y. *et al.* The incidence of endometrial cancer in women with BRCA1 and BRCA2 mutations: An international prospective cohort study. *Gynecol Oncol* (2013).

41. Elloul, S., Kedrin, D., Knoblauch, N.W., Beck, A.H. & Toker, A. The adherens junction protein afadin is an AKT substrate that regulates breast cancer cell migration. *Mol Cancer Res* **12**, 464-76 (2014).

42. Gehrig, P.A. *et al.* Association between uterine serous carcinoma and breast cancer. *Gynecol Oncol* **94**, 208-11 (2004).

43. Belfiore, A. & Frasca, F. IGF and insulin receptor signaling in breast cancer. *J Mammary Gland Biol Neoplasia* **13**, 381-406 (2008).

44. Brennan, K., Offiah, G., McSherry, E.A. & Hopkins, A.M. Tight junctions: a barrier to the initiation and progression of breast cancer? *J Biomed Biotechnol* **2010**, 460607 (2010).

45. DeNardo, D.G. *et al.* Leukocyte complexity predicts breast cancer survival and functionally regulates response to chemotherapy. *Cancer Discov* **1**, 54-67 (2011).

46. Frasca, F. *et al.* The role of insulin receptors and IGF-I receptors in cancer and other diseases. *Arch Physiol Biochem* **114**, 23-37 (2008).

47. Martin, T.A. & Jiang, W.G. Loss of tight junction barrier function and its role in cancer metastasis. *Biochim Biophys Acta* **1788**, 872-91 (2009).

48. Martin, T.A., Watkins, G., Mansel, R.E. & Jiang, W.G. Loss of tight junction plaque molecules in breast cancer tissues is associated with a poor prognosis in patients with breast cancer. *Eur J Cancer* **40**, 2717-25 (2004).

49. Minn, A.J. *et al.* Genes that mediate breast cancer metastasis to lung. *Nature* **436**, 518-24 (2005).

50. Minn, A.J. *et al.* Distinct organ-specific metastatic potential of individual breast cancer cells and primary tumors. *J Clin Invest* **115**, 44-55 (2005).

51. Sachdev, D. Regulation of breast cancer metastasis by IGF signaling. *J Mammary Gland Biol Neoplasia* **13**, 431-41 (2008).

52. Cabioglu, N. *et al.* Chemokine receptors in advanced breast cancer: differential expression in metastatic disease sites with diagnostic and therapeutic implications. *Ann Oncol* **20**, 1013-9 (2009).

53. Hembruff, S.L. & Cheng, N. Chemokine signaling in cancer: Implications on the tumor microenvironment and therapeutic targeting. *Cancer Ther* **7**, 254-267 (2009).

54. Almal, S.H. & Padh, H. Implications of gene copy-number variation in health and diseases. *J Hum Genet* **57**, 6-13 (2012).

55. Bhattacharya, D. & Yusuf, N. Expression of toll-like receptors on breast tumors: taking a toll on tumor microenvironment. *Int J Breast Cancer* **2012**, 716564 (2012).

56. Morrow, P.K. *et al.* Effects of chronic hepatitis C infection on the treatment of breast cancer patients. *Ann Oncol* **21**, 1233-6 (2010).

57. So, E.Y. & Ouchi, T. The application of Toll like receptors for cancer therapy. *Int J Biol Sci* **6**, 675-81 (2010).

58. van Alphen, R.J., Wiemer, E.A., Burger, H. & Eskens, F.A. The spliceosome as target for anticancer treatment. *Br J Cancer* **100**, 228-32 (2009).

59. Couch, F.J. *et al.* Genome-wide association study in BRCA1 mutation carriers identifies novel loci associated with breast and ovarian cancer risk. *PLoS Genet* **9**, e1003212 (2013).

60. Garcia-Closas, M. *et al.* Genome-wide association studies identify four ER negative-specific breast cancer risk loci. *Nat Genet* **45**, 392-8, 398e1-2 (2013).

61. Low, S.K. *et al.* Genome-wide association study of breast cancer in the Japanese population. *PLoS One* **8**, e76463 (2013).

62. Murabito, J.M. *et al.* A genome-wide association study of breast and prostate cancer in the NHLBI's Framingham Heart Study. *BMC Med Genet* **8 Suppl 1**, S6 (2007).

63. Rinella, E.S. *et al.* Genetic variants associated with breast cancer risk for Ashkenazi Jewish women with strong family histories but no identifiable BRCA1/2 mutation. *Hum Genet* **132**, 523-36 (2013).

64. Song, C. *et al.* A genome-wide scan for breast cancer risk haplotypes among African American women. *PLoS One* **8**, e57298 (2013).

65. LA, H. *et al.* A Catalog of Published Genome-Wide Association Studies. Vol. 2013 (National Institute of Health, National Human Genome Research Institute).

66. Jonsdottir, K. *et al.* Validation of expression patterns for nine miRNAs in 204 lymph-node negative breast cancers. *PLoS One* **7**, e48692 (2012).

67. Yoshimoto, N. *et al.* Distinct expressions of microRNAs that directly target estrogen receptor alpha in human breast cancer. *Breast Cancer Res Treat* **130**, 331-9 (2011).

68. Boyle, S.T. & Kochetkova, M. Breast Cancer Stem Cells and the Immune System: Promotion, Evasion and Therapy. *J Mammary Gland Biol Neoplasia* (2014).

69. Cheah, C.Y., Campbell, B.A. & Seymour, J.F. Primary breast lymphoma. *Cancer Treat Rev* (2014).

70. Siriwardana, H.P., Teare, L., Kamel, D. & Inwang, E.R. Toxoplasmosis presenting as a swelling in the axillary tail of the breast and a palpable axillary lymph node mimicking malignancy: a case report. *J Med Case Rep* **5**, 348 (2011).

CHAPTER 7: GENERAL DISCUSSION

Introduction

It is considered that complex disease is associated with changes in gene expression and that any form of variation which causes alteration in gene expression, may be involved in the disease process[12]. The first studies into CNV in humans identified their widespread presence in healthy individuals[10,270] which were confirmed and expanded on in later studies[268,269,274-277,279-281]. An increasing number of reports have since identified CNV's as relevant contributors to human diversity and cancer susceptibility[227,268,270]. Given CNVs encompass more DNA than single nucleotide polymorphisms (SNPs)[273] they are of particular interest as they provide a means to identify regions in the genome where genes associated with disease may reside.

Research into complex diseases aims to discover all variants that predispose individuals to disease development[6]. Breast cancer and colorectal cancer represent two of the most common cancers worldwide, with up to 30% of patients appearing to have a familial origin of disease[217,218,363]. Despite the decades of research into hereditary syndromes associated with breast cancer and colorectal cancers, many patients seeking genetic testing for their condition remain without a molecular diagnosis. It is important for research to continue to uncover the genetic basis of cancer in order to enable better disease prevention, diagnosis and treatment options for patients.

At the commencement of this work, Sanger sequencing and Multiplex Ligation-dependant Probe Amplification (MLPA) were the primary methods for mutation analysis for patients seeking genetic testing for hereditary breast cancer and hereditary colorectal cancers. Both methods are highly targeted and aim to identify (primarily) coding sequence variants (Sanger sequencing) and small duplications or deletions (MLPA) in known cancer susceptibility genes (*BRCA1*, *BRCA2*, *APC*, *MUTYH*, *MLH1*, *MSH2*, *MSH6* and *PMS2*). When no variants can be identified in the genes tested when using these methods, agnostic tests such as genome-wide CNV analysis, can provide a wider perspective of the genome in the search for new genetic causes of disease.

In this thesis, evidence has been provided that supports the hypothesis that CNVs are a potential explanation for a small but significant number of hereditary breast cancer and hereditary colorectal cancer patients who do not harbour germline mutations in genes typically associated with their condition.

General methods and technical limitations

It should be noted that a number of technological limitations became apparent during the course of this work. The Cytogenetic Whole Genome 2.7M (Cyto2.7M) array manufactured by Affymetrix was chosen for use, containing over 400,000 SNP probes and greater than 2.1 million CNV probes (average spacing 1395 bp) and provided the highest genomic coverage of any array at the time this work was commenced. Despite the significant number of SNP probes contained on the array, genotyping information was not made available in the data file and consequently any loss of heterozygosity (LOH) information provided from the array was insufficient for any substantial or reliable analysis. This is a significant drawback of this array as LOH analysis has historically been used for identifying candidate regions associated with disease[364] and could have provided another avenue for data interrogation and deliver additional evidence for the involvement of any CNVs identified in disease.

Furthermore, in order to accurately detect CNVs using arrays, sophisticated algorithms are required. The lack of standardization of the bioinformatics used in CNV analysis has meant data interpretation can be difficult. The Hidden Markov Model (HMM) and Circular Binary Segmentation (CBS) represent the two most common models on which algorithms used in CNV calling software are based and are considered to be the most reliable and accurate[365-367]. It has also been reported that platform specific software perform better in CNV calling compared to platform-independent software algorithms[368] while others have alternatively suggested that the use of multiple algorithms where CNVs being consistently called by different algorithms are most likely real, can increase the reliability of CNV analysis[369]. The proprietary software from Affymetrix, the Chromosome Analysis Suite (ChAS), was solely used for data analysis in this work and employs a HMM-based algorithm that was developed to use with the Cyto2.7M array. As the array data was not compatible for analysis by other user friendly software programs this therefore represents a potential bias and limitation of the CNV analysis.

Overall, genome-wide data revealed a total of 1027 CNVs across 350 samples, with CNVs analysed ranging from 6.03 Kb to 2722.5 Kb in size. The involvement of CNVs smaller than the level of detection (>6 Kb) cannot be ruled out of disease development.

Hereditary breast cancer

As described earlier, it has been suggested that the severity of disease may be explained by the overall burden CNVs place on an individual's genome where increased disease risk is correlated with increasing CNV burden and furthermore that

variation in CNV burden will result in variation in the disease phenotype[282]. No significant difference was detected in the number or size of CNVs between the hereditary breast cancer patients compared to controls suggesting that CNVs (larger than 6 Kb) numerically do not appear to contribute to an increased genomic burden.

A total of 275 rearrangements were identified unique to the hereditary breast cancer cohort which represent candidates for involvement in breast cancer. The generation of large datasets poses a significant obstacle for data analysis and can prove to be problematic when attempting to refine results to a point where one or a few variants, representing the strongest candidates for disease association, may be selected for further investigation.

CNV analysis firstly involved searching for CNVs in and in the vicinity of known cancer susceptibility genes as well as other genes involved in the DSBR pathway (in which *BRCA1* and *BRCA2* reside). Several CNVs were found that disrupted previously reported breast cancer susceptibility genes including *RPA3*, *NBN* (*NBS1*), *MRE11A* and *CYP19A1*[225,370-375] and are likely to have contributed to disease in the affected patients.

Looking more broadly, the genome-wide analysis revealed an increased frequency (>1.55%) of CNVs disrupting *WWOX*. *WWOX* is a well characterized tumour suppressor gene that is known to be associated with breast cancer development, and furthermore, its expression is reported to be associated with the success of tamoxifen treatment[376,377]. Due to ethical constraints, the germline origin of the variants affecting this gene could not be demonstrated, representing an unavoidable limitation of the current study. The frequency of variants detected in this gene suggests that inactivation of *WWOX* may account for a significant proportion of *BRCA1*/*BCRA2* mutation negative hereditary breast cancer patients.

Looking specifically at rare genes disrupted by a CNV, which has been reported to assist in the identification of highly penetrant disease factors[285], 78 genes were revealed in 27 patients. Of particular interest was the identification of several genes previously implicated in cancer, including breast cancer[368,369], of which a germline intronic deletion was identified in the tumour suppressor *FHIT*. Intronic deletions in *FHIT* have previously been implicated in pancreatic cancer[378]. The intronic variant identified by the current study is similarly proposed to disrupt *FHIT* expression and result in disease. In relation to breast cancer, *FHIT* is known to be genetically and epigenetically modified in tumours[379-384]. *FHIT* expression is also considered to be

protective against *HER2*-driven breast tumour development[385], while reduced *FHIT* expression is associated with poor prognosis[386]. Overall, it appears that analysis of rare genes can aid in the identification of potentially causative variants contributing to disease.

HNPCC

An increased CNV burden was not observed among the HNPCC patients either, however a decreased mean size of CNVs was detected in HNPCC patients compared to controls (p=0.0165). In a previous publication looking at both MMR mutation negative and MMR mutation positive patients we observed conversely an increased average size of CNVs in patients which suggested that this was related to an increased genomic burden[387]. These discrepancies may be attributed to the inequity of sample populations between studies (the current compared 125 patients and 40 controls whereas the previous compared 96 patients and 384 controls[387]), the limited number of controls included in the current work, the type of array chosen (noting differences in both the array coverage and density), as well as the CNV calling algorithm used by the different analysis software's[365-367].

The CNV analysis of the HNPCC cohort was relatively unremarkable compared to that of the FAP and the hereditary breast cancer cohorts, revealing no obvious CNVs that could account for disease. This may not be surprizing since tumours arising in a setting of LS tend not to have any significant LOH, rather they display an MSI phenotype. It is speculated that there could be a relationship between LS and a reduced risk of acquiring a CNV and the mechanism for this may be associated with the triggering of other DNA repair pathways, e.g. DSBR in the absence of MMR. Overall it appears that the mechanisms for disease development in HNPCC are more elusive than either hereditary breast cancer of FAP for the patient cohorts tested.

With the help of bioinformatic analysis tools, such as pathway analysis software and miR annotation databases, more in-depth investigations may be undertaken to uncover more complex associations with disease[359-361]. For example, pathway analysis of the 300 plus genes uniquely identified in the HNPCC cohort suggested the enrichment of pathways involved in metabolism. Metabolic processes are well characterized in association with carcinogenesis[388,389] and therefore it is likely that one or more of the 21 genes identified driving the enrichment of this pathway may contribute to colorectal cancer risk in the affected individuals. Overall, it appears that genome-wide CNV analysis can be problematic when large datasets fail to reveal obvious regions of

interest. In order to gain insight into such datasets bioinformatic tools such as pathway analysis and miR annotation can prove useful.

To further elucidate the contribution of CNVs in LS, investigation was also undertaken with the aim of detecting small duplications and deletions which may reside deep in the introns of *MLH1*, *MSH2* and *MSH6.* Aberrations harboured within intronic regions of genes have emerged as a cause of gene inactivation that may give rise to disease[350-353]. Several studies have shown that mutations in deep intronic regions of known cancer susceptibility genes contribute to the pathogenesis of hereditary breast cancer and FAP[350,353,390], however less is understood of their overall prevalence in HNPCC. Unfortunately no deleterious intronic variants were identified in *MLH1*, *MSH2* and *MSH6* in any of the HNPCC patients screened suggesting that these deep intronic variants are rare (<1%). A single study investigating intronic variants resulting in aberrant splicing in FAP has reported that deep intronic variants account for up to 8% (10 in 125 patients) of *APC* mutation negative FAP patients[350]. The difference in the frequency of deep intronic mutations observed between these hereditary colorectal cancers is considered to be statistically significant (χ^2=3.9086, *p=0.04804*). Evidence provided in this work does not support the frequent involvement of deep intronic variants in contribution to LS.

FAP

Like the analysis of the hereditary breast cancer cohort, no significant differences were detected in the number or size of CNVs between the polyposis patients compared to the controls. These findings suggest that CNVs larger than 6 Kb numerically do not appear to contribute to an increased genomic burden in polyposis either.

Genome-wide data can enhance a targeted analysis approach as it enables the examination of genomic regions not just within, but also in the vicinity of known cancer susceptibility genes. Consequently analysis can encompass gene regulatory regions e.g. promoter sequences or CpG islands not always routinely tested in a clinical setting, and whereby variants harboured in these regions may also contribute to the development of disease. An example of this was the identification of a 31 Kb CN loss that was located directly within the promoter 1B region of *APC* in one FAP patient submitted for genetic testing in 1998. Until recently, current diagnostic testing using Sanger sequencing and MLPA would have failed to detect this likely causative rearrangement. In 2011, Rohlin *et al*.[112] reported the first evidence of promoter 1B involvement in FAP and since then this region has been incorporated in to MLPA kits

used by diagnostic laboratories testing this condition. The results presented in the current body of work provide further support for and demonstrates the benefits of including non-coding genomic regions (in general) into diagnostic screening.

Overall the genome-wide analysis revealed 142 CNVs unique to the polyposis cohort which represents candidate regions for involvement in polyposis. Of particular interest was a CN loss located on chromosome 18 at 18p11.32 that affected nearly 9% of the patients screened. Independent studies have reported CNVs in this region to have potential associations with disease including colorectal cancer[391-395]. This region was found to harbour a master-regulatory element lnc-RNA. Lnc-RNA functions include post-transcriptional regulation of gene expression, regulation of epigenetic marks, gene activation in *cis* and have been shown to influence processes such as pluripotency[396-398]. Lnc-RNAs have been identified to contribute to colorectal cancer (see[399] for a recent review). Most recently Ma *et al.*[400] has reported the existence of a novel lnc-RNA, *CCAL,* that is believed to be an oncogenic regulator in colorectal cancer tumorigenesis. Furthermore, it is reported that high expression of *CCAL* in tumours is associated with shorter patient survival and a less favourable outcomes to chemotherapy treatments[400]. It is believed that *CCAL* promotes disease progression via targeting *AP-2*α which results in the activation of WNT signalling[400]. Additional evidence is provided in the current body of work which suggests the possible role of other lnc-RNAs in the aetiology of colorectal cancer.

Similarities and differences

Patients conforming to hereditary colorectal cancer syndromes such as HNPCC are reported to have an increased risk of developing breast cancer[26,27,401-404] though reports disputing the existence of any such association have also been published (recently reviewed in[28]). While investigating this association was not a specific aim during this work, data is provided that when taken together supports this theory.

Overall, several genes disrupted by a CNV (*ARHGD1B*, *B2M*, *DCDC1*, *GPR128*, *IMMP2L*, *NAMPT*, *TFG* and *TRIM69*) were revealed in-common to all three patient cohorts and were not featured among the genes disrupted by a CNV in any of the control genomes. Of these genes, *NAMPT* has previously been implicated in both breast cancer and colorectal cancer[405-408] while *B2M* has been associated with just colorectal cancer[409]. These results imply that since these genes have been observed to be disrupted by a CNV across all three cancer types, that they may either be involved in common mechanisms of cancer development, or alternatively that they may

associated with a subset of patients who specifically have a heighted risk of developing both breast cancer and colorectal cancer, and not just one or the other.

It was also observed that the hereditary breast cancer cohort had more genes disrupted by a CNV that were in-common with genes also disrupted by a CNV in the FAP cohort (18%) compared to the HNPCC cohort (9%). Furthermore, genes disrupted by a CNV in the FAP cohort also showed a higher affinity to those disrupted by a CNV in the hereditary breast cancer cohort (17%) than the other colorectal cancer syndrome HNPCC (13%). A possible explanation for this could again be related to the fact that tumours associated with LS tend to harbour MSI rather than LOH, while in tumours associated with FAP and breast cancers LOH is commonly observed.

General conclusions

Hereditary breast cancer and hereditary colorectal cancers are complex diseases for which the underlying genetic basis is continuing to emerge. Genome-wide CN analysis provides a valuable tool for identifying novel genomic regions which may underpin disease development in patients. It appears that overall, CNV do not intrinsically contribute to an increased genomic burden that is associated with increased risk of disease in either hereditary breast cancer or hereditary colorectal cancers. Several regions of significant interest were revealed from the genome-wide analyses that are considered likely contributors to disease development in the affected individuals. This has included the identification of CNVs disrupting *WWOX* and *FHIT* in three unrelated hereditary breast cancer patients, CNVs in *APC*, *DCC*, *MLH1* and *CTNNB1I* among four unrelated FAP patients, in addition to five other FAP patients whom were identified to harbour CNV losses at 18p11.32. In this body of work, evidence has also been provided that reinforces the benefits in using bioinformatic analysis tools including pathway analysis and miR annotation when genome-wide analysis yields many disease candidate targets to follow-up. Pathway analysis of the 317 genes uniquely identified in the HNPCC cohort revealed enrichment among pathways involved in metabolism and these have been known to be required for cancer development. It is likely that the affected loci driving these associations may contribute to colorectal cancer risk in the affected individuals.

Overall evidence has been provided in this thesis which supports the involvement of CNVs in a small but significant number of hereditary breast cancer and hereditary colorectal cancer patients. Further investigation are suggested which may continue to uncover the mechanisms involved in these diseases.

Future directions

While the evidence provided in this body of work represents a significant contribution to our understanding of the role of CNVs in hereditary breast cancer and hereditary colorectal cancers, ongoing research is needed to fully elucidate the genetic basis of these conditions. In particular several areas have been identified in which research should be focused, including:

1. ***Improving CNV array designs*** to enable reliable detection of both CNVs and regions of LOH and therefore provide a holistic data source for the detection of regions potentially associated with disease;
2. ***Standardizing the bioinformatics behind CNV calling*** and therefore improve CNV data interpretation and ensure minimal false-positive calls are carried through for further investigation;
3. ***Investigating the contribution of CNVs <6 Kb in the development of hereditary breast cancer and hereditary colorectal cancers***. As CNV analysis contained in the current body of work was limited to the detection of CNVs greater than this, smaller CNVs may still remain a significant cause of disease in these patients.
4. ***Undertaking functional studies into the rearrangements identified*** e.g. the CNVs revealed in *WWOX*, *FHIT* and 18p11.32, to gain insight into the mechanisms by which they support disease development;
5. ***Conducting segregation analysis*** that involves testing both affected and unaffected family members to demonstrate the transference of identified variants from one generation to the next, while furthermore providing stronger evidence for their involvement in disease;
6. ***Expand the current study to include more samples, particularly controls***. This will provide greater power to detect significant relationships potentially revealed in association testing and would help overcome potential biases created by small sample sizes. In the context of this work, this would be beneficial in confirming: (1) the significant decreased mean size of CNVs detected in HNPCC patients compared to the controls, and (2) the frequency of CN losses observed at 18p11.32 in FAP that were detected in 9% of the patients screened.
7. ***Investigating the reasons behind why a lack of remarkable CNVs was identified in the HNPCC cohort.*** No CNVs were revealed in any of the other 18 genes in the MMR pathway or other colorectal cancer susceptibility genes however in contrast many regions of interest were identified in the analysis of

the hereditary breast cancer and FAP cohorts. This may include the testing of the proposed theory that mechanisms associated with failed MMR and the accumulation of MSI in some way protects against the acquisition of LOH.

8. ***Improving the bioinformatics behind pathway analysis and miR annotation programs***, including expanding research into the genetic networks on which these software's are based, therefore provide a more reliable and a more accurate analysis. This will hopefully one day be expanded to include additional resources that incorporate gene-environmental interactions, as well as other genomic factors such as linc-RNAs.
9. ***Conducting a larger investigation into the contribution of deep intronic variants in the development of LS.*** This should also include investigation into the observed absence of naturally occurring splice variants in HNPCC patient cell lines, focusing on the possible impact of Epstein Barr virus (EBV)-immortalization compared to blood-derived RNA sources for fragment analysis.
10. ***Undertaking a larger case-control study involving Next-Gen Sequencing and more CNV analysis*** to further tease out the contribution of CNVs in disease development (particularly those <6 Kb). RNA Seq could be helpful in demonstrating the functional impact of CNVs e.g. when they result in aberrant gene transcripts and changes in gene expression. Furthermore, Methyl Seq could provide insight into the underlying epigenetic landscape, which may also be altered in the presence of CNVs and another possible explanation for disease.

APPENDICES

List of abbreviations

BER	Base Excision Repair
BMI	Body mass index
bp	Base pairs
CBS	circular binary segmentation
ChAS	Chromosome Analysis Suite
CN	Copy number
CNV	Copy number variation
COSMIC	Catalogue of somatic mutations in cancer
Cyto2.7M	Affymetrix Cytogenetic Whole Genome 2.7M Microarray
DCIS	Ductal carcinoma in-situ
DGV	Database of Genomic Variants
DNA	Deoxyribose nucleic acid
DSB	Double-strand breaks
DSBR	Double-strand break repair
DR	Direct Reverse
EBV	Epstein Barr Virus
ER	Oestrogen receptor
FAP	Familial adenomatous polyposis
GI	Gastrointestinal
HER2	human epidermal growth factor receptor 2
HMM	Hidden Markov Model
HNPCC	Hereditary non-polyposis colorectal cancer

HR	Homologous Recombination
ERT	Endocrine Replacement Therapy
IMPS	Hereditary mixed polyposis syndrome
InSiGHT	International Society for Gastrointestinal Hereditary Tumours
Kb	Kilo base
JP	Juvenile polyposis
LCIS	Lobular carcinoma in-situ
LOH	Loss of heterozygosity
LS	Lynch syndrome
MAP	*MUTYH* associated polyposis
Mb	Mega base
MCR	Mutation cluster region
MiR	Micro (size) RNA
MMEJ	Microhomoloy mediated end-joining
MMR	Mismatch repair
MLPA	Multiplex ligation-dependant probe amplification
MRI	Magnetic resonance imaging
mRNA	Messenger RNA
MSI	Microsatellite instability
NER	Nucleotide excision repair
NHEJ	Non-homologous end-joining
PJS	Peutz Jeghers syndrome
PR	Progesterone receptor
RNA	Ribose nucleic acid

ROS	Reactive oxygen species
SNP	Single nucleotide polymorphism
TCGA	The Cancer Genome Atlas
TNBC	Triple negative breast cancer
UV	Ultra-violet (light)

BIBLIOGRAPHY

1. Talseth-Palmer, B.A. & Scott, R.J. Genetic Variation and its Role in Malignancy. *International journal of Biomedical science* **7**, 14 (2011).

2. Gejman, P.V., Sanders, A.R. & Duan, J. The Role of Genetics in the Etiology of Schizophrenia. *Psychiatr Clin North Am* **33**, 31 (2010).

3. Gregersen, P.K. & Olsson, L.M. Recent Advances in the Genetics of Autoimmune Disease. *Ann Rev Immunology* **27**, 28 (2009).

4. Thanassoulis, G., Ramachandran, S. & Vasan, M.D. Genetic Cardiovascular Risk Prediction - Will we get there? *Circulation* **122**, 11 (2010).

5. Ripattia, S. *et al.* A multilocus genetic risk score for coronary heart disease: case control and prospective cohort analyses. *Lancet* **376**, 7 (2010).

6. Shlien, A. & Malkin, D. Copy number variations and cancer. *Genome Med* **1**, 62 (2009).

7. Escaramis, G., Docampo, E. & Rabionet, R. A decade of structural variants: description, history and methods to detect structural variation. *Brief Funct Genomics* (2015).

8. Bridges, C.B., Skoog, E.N. & Li, J.C. Genetical and Cytological Studies of a Deficiency (Notopleural) in the Second Chromosome of Drosophila Melanogaster. *Genetics* **21**, 788-95 (1936).

9. Dhami, P. *et al.* Exon array CGH: detection of copy-number changes at the resolution of individual exons in the human genome. *Am J Hum Genet* **76**, 750-62 (2005).

10. Iafrate, A.J. *et al.* Detection of large-scale variation in the human genome. *Nat Genet* **36**, 949-51 (2004).

11. Stankiewicz, P. & Lupski, J.R. Molecular-evolutionary mechanisms for genomic disorders. *Curr Opin Genet Dev* **12**, 312-9 (2002).

12. Shelling, A.N. & Ferguson, L.R. Genetic variation in human disease and a new role for copy number variants. *Mutat Res* **622**, 33-41 (2007).

13. Nordling, C.O. [Theories and statistics of cancer.]. *Nord Med* **47**, 817-20 (1952).

14. Han, J., Colditz, G.A., Liu, J.S. & Hunter, D.J. Genetic variation in XPD, sun exposure, and risk of skin cancer. *Cancer Epidemiol Biomarkers Prev* **14**, 1539-44 (2005).

15. Boyle, P. Cancer, cigarette smoking and premature death in Europe: a review including the Recommendations of European Cancer Experts Consensus Meeting, Helsinki, October 1996. *Lung Cancer* **17**, 1-60 (1997).

16. Milanowska, K. *et al.* REPAIRtoire--a database of DNA repair pathways. *Nucleic Acids Res* **39**, D788-92 (2011).

17. Larrea, A.A., Lujan, S.A. & Kunkel, T.A. SnapShot: DNA mismatch repair. *Cell* **141**, 730 e1 (2010).

18. Wyman, C., Ristic, D. & Kanaar, R. Homologous recombination-mediated double-strand break repair. *DNA Repair (Amst)* **3**, 827-33 (2004).

19. Torre, L.A. *et al.* Global cancer statistics, 2012. *CA Cancer J Clin* **65**, 87-108 (2015).

20. Lichtenstein, P. *et al.* Environmental and heritable factors in the causation of cancer--analyses of cohorts of twins from Sweden, Denmark, and Finland. *N Engl J Med* **343**, 78-85 (2000).

21. Lynch, H.T. & de la Chapelle, A. Hereditary colorectal cancer. *N Engl J Med* **348**, 919-32 (2003).

22. Gatalica, Z. & Torlakovic, E. Pathology of the hereditary colorectal carcinoma. *Fam Cancer* **7**, 15-26 (2008).

23. Sasada, T. *et al.* Chlorinated Water Modulates the Development of Colorectal Tumors with Chromosomal Instability and Gut Microbiota in Apc-Deficient Mice. *PLoS One* **10**, e0132435 (2015).

24. Aaltonen, L.A. *et al.* Incidence of hereditary nonpolyposis colorectal cancer and the feasibility of molecular screening for the disease. *N Engl J Med* **338**, 1481-7 (1998).

25. Aaltonen, L., Johns, L., Jarvinen, H., Mecklin, J.P. & Houlston, R. Explaining the familial colorectal cancer risk associated with mismatch repair (MMR)-deficient and MMR-stable tumors. *Clin Cancer Res* **13**, 356-61 (2007).

26. Scott, R.J. & Ashton, K.A. Familial breast and bowel cancer: does it exist? *Hered Cancer Clin Pract* **2**, 25-9 (2004).

27. Scott, R.J. *et al.* Hereditary nonpolyposis colorectal cancer in 95 families: differences and similarities between mutation-positive and mutation-negative kindreds. *Am J Hum Genet* **68**, 118-127 (2001).

28. Win, A.K., Lindor, N.M. & Jenkins, M.A. Risk of breast cancer in Lynch syndrome: a systematic review. *Breast Cancer Res* **15**, R27 (2013).

29. Hou, N. *et al.* Hormone replacement therapy and breast cancer: heterogeneous risks by race, weight, and breast density. *J Natl Cancer Inst* **105**, 1365-72 (2013).

30. Pearson, J.P. & Brownlee, I.A. The interaction of large bowel microflora with the colonic mucus barrier. *Int J Inflam* **2010**, 321426 (2010).

31. Gibson, G.R. & Roberfroid, M.B. Dietary modulation of the human colonic microbiota: introducing the concept of prebiotics. *J Nutr* **125**, 1401-12 (1995).

32. Guarner, F. & Malagelada, J.R. Gut flora in health and disease. *Lancet* **361**, 512-9 (2003).

33. Tuohy, K.M., Kolida, S. & Gibson, G.R. Use of probiotics and prebiotics for improving gut health. *Agro Food Industry Hi-Tech* **15**, 2 (2004).

34. Ebersbach, T. *et al.* Certain dietary carbohydrates promote Listeria infection in a guinea pig model, while others prevent it. *Int J Food Microbiol* **140**, 218-24 (2010).

35. Licht, T.R. *et al.* Effects of apples and specific apple components on the cecal environment of conventional rats: role of apple pectin. *BMC Microbiol* **10**, 13 (2010).

36. Petersen, A. *et al.* Some putative prebiotics increase the severity of Salmonella enterica serovar Typhimurium infection in mice. *BMC Microbiol* **9**, 245 (2009).

37. Gueimonde, M., Ouwehand, A., Huhtinen, H., Salminen, E. & Salminen, S. Qualitative and quantitative analyses of the bifidobacterial microbiota in the colonic mucosa of patients with colorectal cancer, diverticulitis and inflammatory bowel disease. *World J Gastroenterol* **13**, 3985-9 (2007).

38. Mai, V., McCrary, Q.M., Sinha, R. & Glei, M. Associations between dietary habits and body mass index with gut microbiota composition and fecal water genotoxicity: an observational study in African American and Caucasian American volunteers. *Nutr J* **8**, 49 (2009).

39. O'Keefe, S.J. Nutrition and colonic health: the critical role of the microbiota. *Curr Opin Gastroenterol* **24**, 51-8 (2008).

40. O'Keefe, S.J. *et al.* Products of the colonic microbiota mediate the effects of diet on colon cancer risk. *J Nutr* **139**, 2044-8 (2009).

41. Pryde, S.E., Duncan, S.H., Hold, G.L., Stewart, C.S. & Flint, H.J. The microbiology of butyrate formation in the human colon. *FEMS Microbiol Lett* **217**, 133-9 (2002).

42. Cummings, J.H., Bingham, S.A., Heaton, K.W. & Eastwood, M.A. Fecal weight, colon cancer risk, and dietary intake of nonstarch polysaccharides (dietary fiber). *Gastroenterology* **103**, 1783-9 (1992).

43. Guerin, A. *et al.* Risk of developing colorectal cancer and benign colorectal neoplasm in patients with chronic constipation. *Aliment Pharmacol Ther* **40**, 83-92 (2014).

44. Erhardt, J.G., Lim, S.S., Bode, J.C. & Bode, C. A diet rich in fat and poor in dietary fiber increases the in vitro formation of reactive oxygen species in human feces. *J Nutr* **127**, 706-9 (1997).

45. Kuhajda, F.P. Fatty-acid synthase and human cancer: new perspectives on its role in tumor biology. *Nutrition* **16**, 202-8 (2000).

46. Kuhajda, F.P. Fatty acid synthase and cancer: new application of an old pathway. *Cancer Res* **66**, 5977-80 (2006).

47. Menendez, J.A. & Lupu, R. Mediterranean dietary traditions for the molecular treatment of human cancer: anti-oncogenic actions of the main olive oil's monounsaturated fatty acid oleic acid (18:1n-9). *Curr Pharm Biotechnol* **7**, 495-502 (2006).

48. Menendez, J.A. & Lupu, R. Oncogenic properties of the endogenous fatty acid metabolism: molecular pathology of fatty acid synthase in cancer cells. *Curr Opin Clin Nutr Metab Care* **9**, 346-57 (2006).

49. Menendez, J.A. & Lupu, R. Fatty acid synthase and the lipogenic phenotype in cancer pathogenesis. *Nat Rev Cancer* **7**, 763-77 (2007).

50. Osborne, N.J. *et al.* HFE C282Y homozygotes are at increased risk of breast and colorectal cancer. *Hepatology* **51**, 1311-8 (2010).

51. Brookes, M.J. *et al.* A role for iron in Wnt signalling. *Oncogene* **27**, 966-75 (2008).

52. Jasperson, K.W., Tuohy, T.M., Neklason, D.W. & Burt, R.W. Hereditary and familial colon cancer. *Gastroenterology* **138**, 2044-58 (2010).

53. Chua, A.C., Klopcic, B., Lawrance, I.C., Olynyk, J.K. & Trinder, D. Iron: an emerging factor in colorectal carcinogenesis. *World J Gastroenterol* **16**, 663-72 (2010).

54. Franco, A., Sikalidis, A.K. & Solis Herruzo, J.A. Colorectal cancer: influence of diet and lifestyle factors. *Rev Esp Enferm Dig* **97**, 432-48 (2005).

55. Giovannucci, E. Diet, body weight, and colorectal cancer: a summary of the epidemiologic evidence. *J Womens Health (Larchmt)* **12**, 173-82 (2003).

56. Giovannucci, E. & Goldin, B. The role of fat, fatty acids, and total energy intake in the etiology of human colon cancer. *Am J Clin Nutr* **66**, 1564S-1571S (1997).

57. Howe, G.R. *et al.* The relationship between dietary fat intake and risk of colorectal cancer: evidence from the combined analysis of 13 case-control studies. *Cancer Causes Control* **8**, 215-28 (1997).

58. Kushi, L. & Giovannucci, E. Dietary fat and cancer. *Am J Med* **113 Suppl 9B**, 63S-70S (2002).

59. Slattery, M.L., Potter, J.D., Duncan, D.M. & Berry, T.D. Dietary fats and colon cancer: assessment of risk associated with specific fatty acids. *Int J Cancer* **73**, 670-7 (1997).

60. Vinikoor, L.C. *et al.* trans-Fatty acid consumption and its association with distal colorectal cancer in the North Carolina Colon Cancer Study II. *Cancer Causes Control* **21**, 171-80 (2010).

61. Tomey, K.M. *et al.* Dietary fat subgroups, zinc, and vegetable components are related to urine F2a-isoprostane concentration, a measure of oxidative stress, in midlife women. *J Nutr* **137**, 2412-9 (2007).

62. Tamura, K. *et al.* Mechanism of carcinogenesis in familial tumors. *Int J Clin Oncol* **9**, 232-45 (2004).

63. Bronner, C.E. *et al.* Mutation in the DNA mismatch repair gene homologue hMLH1 is associated with hereditary non-polyposis colon cancer. *Nature* **368**, 258-61 (1994).

64. Fishel, R. *et al.* The human mutator gene homolog MSH2 and its association with hereditary nonpolyposis colon cancer. *Cell* **75**, 1027-38 (1993).

65. Leach, F.S. *et al.* Mutations of a mutS homolog in hereditary nonpolyposis colorectal cancer. *Cell* **75**, 1215-25 (1993).

66. Lindblom, A., Tannergard, P., Werelius, B. & Nordenskjold, M. Genetic mapping of a second locus predisposing to hereditary non-polyposis colon cancer. *Nat Genet* **5**, 279-82 (1993).

67. Papadopoulos, N. *et al.* Mutation of a mutL homolog in hereditary colon cancer. *Science* **263**, 1625-9 (1994).

68. Peltomaki, P. *et al.* Genetic mapping of a locus predisposing to human colorectal cancer. *Science* **260**, 810-2 (1993).

69. Tannergard, P., Zabarovsky, E., Stanbridge, E., Nordenskjold, M. & Lindblom, A. Sublocalization of a locus at 3p21.3-23 predisposing to hereditary nonpolyposis colon cancer. *Hum Genet* **94**, 210-4 (1994).

70. Lynch, H.T. *et al.* Lynch Syndrome-Associated Extracolonic Tumors Are Rare in Two Extended Families With the Same EPCAM Deletion. *Am J Gastroenterol* (2011).

71. Kuiper, R.P. *et al.* Recurrence and variability of germline EPCAM deletions in Lynch syndrome. *Hum Mutat* **32**, 407-14 (2011).

72. Kempers, M.J. *et al.* Risk of colorectal and endometrial cancers in EPCAM deletion-positive Lynch syndrome: a cohort study. *Lancet Oncol* **12**, 49-55 (2011).

73. Lynch, H.T. & de la Chapelle, A. Genetic susceptibility to non-polyposis colorectal cancer. *J Med Genet* **36**, 801-18 (1999).

74. Bhaijee, F. & Brown, A.S. Muir-Torre syndrome. *Arch Pathol Lab Med* **138**, 1685-9 (2014).

75. Hamilton, S.R. *et al.* The molecular basis of Turcot's syndrome. *N Engl J Med* **332**, 839-47 (1995).

76. Galiatsatos, P. & Foulkes, W.D. Familial adenomatous polyposis. *Am J Gastroenterol* **101**, 385-98 (2006).

77. Woodage, T. *et al.* The APCI1307K allele and cancer risk in a community-based study of Ashkenazi Jews. *Nat Genet* **20**, 62-5 (1998).

78. Laken, S.J. *et al.* Familial colorectal cancer in Ashkenazim due to a hypermutable tract in APC. *Nat Genet* **17**, 79-83 (1997).

79. Lipton, L. & Tomlinson, I. The genetics of FAP and FAP-like syndromes. *Fam Cancer* **5**, 221-6 (2006).

80. Morak, M., Laner, A., Bacher, U., Keiling, C. & Holinski-Feder, E. MUTYH-associated polyposis - variability of the clinical phenotype in patients with biallelic and monoallelic MUTYH mutations and report on novel mutations. *Clin Genet* **78**, 353-63 (2010).

81. Molatore, S. *et al.* MUTYH mutations associated with familial adenomatous polyposis: functional characterization by a mammalian cell-based assay. *Hum Mutat* **31**, 159-66 (2010).

82. Lubbe, S.J., Di Bernardo, M.C., Chandler, I.P. & Houlston, R.S. Clinical implications of the colorectal cancer risk associated with MUTYH mutation. *J Clin Oncol* **27**, 3975-80 (2009).

83. Ali, M. *et al.* Characterization of mutant MUTYH proteins associated with familial colorectal cancer. *Gastroenterology* **135**, 499-507 (2008).

84. van Hattem, W.A. *et al.* Large genomic deletions of SMAD4, BMPR1A and PTEN in juvenile polyposis. *Gut* **57**, 623-7 (2008).

85. Huang, S.C. *et al.* Genetic heterogeneity in familial juvenile polyposis. *Cancer Res* **60**, 6882-5 (2000).

86. Gammon, A., Jasperson, K., Kohlmann, W. & Burt, R.W. Hamartomatous polyposis syndromes. *Best Pract Res Clin Gastroenterol* **23**, 219-31 (2009).

87. Le Meur, N. *et al.* Complete germline deletion of the STK11 gene in a family with Peutz-Jeghers syndrome. *Eur J Hum Genet* **12**, 415-8 (2004).

88. Schumacher, V. *et al.* STK11 genotyping and cancer risk in Peutz-Jeghers syndrome. *J Med Genet* **42**, 428-35 (2005).

89. Rustgi, A.K. The genetics of hereditary colon cancer. *Genes Dev* **21**, 2525-38 (2007).

90. Half, E., Bercovich, D. & Rozen, P. Familial adenomatous polyposis. *Orphanet J Rare Dis* **4**, 22 (2009).

91. Berk, T., Cohen, Z. & Cullen, J.B. Familial polyposis and the role of the preventive registry. *Can Med Assoc J* **124**, 1427-8 (1981).

92. Stenson, P.D. *et al.* The Human Gene Mutation Database: 2008 update. *Genome Med* **1**, 13 (2009).

93. Al-Tassan, N. *et al.* Inherited variants of MYH associated with somatic G:C-->T:A mutations in colorectal tumors. *Nat Genet* **30**, 227-32 (2002).

94. Pezzi, A. *et al.* Relative role of APC and MUTYH mutations in the pathogenesis of familial adenomatous polyposis. *Scand J Gastroenterol* **44**, 1092-100 (2009).

95. Claes, K. *et al.* The genetics of familial adenomatous polyposis (FAP) and MutYH-associated polyposis (MAP). *Acta Gastroenterol Belg* **74**, 421-6 (2011).

96. Brand, R., Nielsen, M., Lynch, H. & Infante, E. MUTYH-Associated Polyposis. in *GeneReviews(R)* (eds. Pagon, R.A. *et al.*) (Seattle (WA), 2013).

97. Scates, D., Clarke, S., Phillips, R. & Venitts, S. Lack of telomeres in desmoids occurring sproadically and in association with Familial Adenomatous Polyposis. *Br J Surg* **85**, 4 (1998).

98. Bowden, N.A., Croft, A. & Scott, R.J. Gene expression profiling in Familial Adenomatous Polyposis adenomas and desmoid disease. *Hered Cancer Clin Pract* **5**, 18 (2007).

99. Knudsen, A.L., Bisgaard, M.L. & Bulow, S. Attenuated familial adenomatous polyposis (AFAP). A review of the literature. *Familial Cancer* **2**, 12 (2003).

100. Schouten, J.P. *et al.* Relative quantification of 40 nucleic acid sequences by multiplex ligation-dependent probe amplification. *Nucleic Acids Res* **30**, e57 (2002).

101. Segditsas, S. & Tomlinson, I. Colorectal cancer and genetic alterations in the Wnt pathway. *Oncogene* **25**, 7531-7 (2006).

102. Polakis, P. Mutations in the APC gene and their implications for protein structure and function. *Curr Opin Genet Dev* **5**, 66-71 (1995).

103. Fearon, E.R. Molecular genetics of colorectal cancer. *Annu Rev Pathol* **6**, 479-507 (2011).

104. Aretz, S. *et al.* Large submicroscopic genomic APC deletions are a common cause of typical familial adenomatous polyposis. *J Med Genet* **42**, 185-92 (2005).

105. Gismondi, V. *et al.* 310 basepair APC deletion with duplication of breakpoint (439ins15del310) in an Italian polyposis patient. *Hum Mutat* **Suppl 1**, S220-2 (1998).

106. Charames, G.S. *et al.* A large novel deletion in the APC promoter region causes gene silencing and leads to classical familial adenomatous polyposis in a Manitoba Mennonite kindred. *Hum Genet* **124**, 535-41 (2008).

107. Segditsas, S. *et al.* Promoter hypermethylation leads to decreased APC mRNA expression in familial polyposis and sporadic colorectal tumours, but does not substitute for truncating mutations. *Exp Mol Pathol* **85**, 201-6 (2008).

108. Tsuchiya, T. *et al.* Distinct methylation patterns of two APC gene promoters in normal and cancerous gastric epithelia. *Oncogene* **19**, 3642-6 (2000).

109. De Rosa, M. *et al.* Alternative splicing and nonsense-mediated mRNA decay in the regulation of a new adenomatous polyposis coli transcript. *Gene* **395**, 8-14 (2007).

110. Lambertz, S. & Ballhausen, W.G. Identification of an alternative 5' untranslated region of the adenomatous polyposis coli gene. *Hum Genet* **90**, 650-2 (1993).

111. Hosoya, K. *et al.* Adenomatous polyposis coli 1A is likely to be methylated as a passenger in human gastric carcinogenesis. *Cancer Lett* **285**, 182-9 (2009).

112. Rohlin, A. *et al.* Inactivation of promoter 1B of APC causes partial gene silencing: evidence for a significant role of the promoter in regulation and causative of familial adenomatous polyposis. *Oncogene* (2011).

113. Horii, A., Nakatsuru, S., Ichii, S., Nagase, H. & Nakamura, Y. Multiple forms of the APC gene transcripts and their tissue-specific expression. *Hum Mol Genet* **2**, 283-7 (1993).

114. Yamaguchi, S. *et al.* MUTYH-associated colorectal cancer and adenomatous polyposis. *Surg Today* **44**, 593-600 (2014).

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

116. Aretz, S. *et al.* Somatic APC mosaicism: a frequent cause of familial adenomatous polyposis (FAP). *Hum Mutat* **28**, 985-92 (2007).

117. Baert-Desurmont, S. *et al.* A remarkable APC mosaicism with two mutant alleles in a family with familial adenomatous polyposis. *Am J Med Genet A* **155A**, 1500-2 (2011).

118. Davidson, S., Leshanski, L., Rennert, G., Eidelman, S. & Amikam, D. Maternal mosaicism for a second mutational event--a novel deletion--in a familial adenomatous polyposis family harboring a new germ-line mutation in the alternatively spliced-exon 9 region of APC. *Hum Mutat* **19**, 83-4 (2002).

119. Hes, F.J. *et al.* Somatic APC mosaicism: an underestimated cause of polyposis coli. *Gut* **57**, 71-6 (2008).

120. Necker, J., Kovac, M., Attenhofer, M., Reichlin, B. & Heinimann, K. Detection of APC germ line mosaicism in patients with de novo familial adenomatous polyposis: a plea for the protein truncation test. *J Med Genet* **48**, 526-9 (2011).

121. Anastas, J.N. & Moon, R.T. WNT signalling pathways as therapeutic targets in cancer. *Nat Rev Cancer* **13**, 11-26 (2013).

122. Polakis, P. Wnt signaling in cancer. *Cold Spring Harb Perspect Biol* **4**(2012).

123. Nusse, R. & Varmus, H.E. Wnt genes. *Cell* **69**, 1073-87 (1992).

124. Munemitsu, S., Albert, I., Souza, B., Rubinfeld, B. & Polakis, P. Regulation of intracellular beta-catenin levels by the adenomatous polyposis coli (APC) tumor-suppressor protein. *Proc Natl Acad Sci U S A* **92**, 3046-50 (1995).

125. Huelsken, J. & Behrens, J. The Wnt signalling pathway. *J Cell Sci* **115**, 3977-8 (2002).

126. Maise, K., Li, F., Chong, Z.Z. & Chen, S.Y. The WNT Signalling Pathway: aging gracefully as a protectionist? *Pharmacol Ther* **118**, 23 (2008).

127. MacDonald, B.T., Tamai, K. & He, X. WNT/b-Catenin Signalling: Components, Mechanisms, and Disease. *Dev Cell* **17**, 17 (2009).

128. Obermair, A. *et al.* Risk of Endometrial Cancer for women diagnosed with HNPCC-related colorectal cancer. *International journal of cancer* **127**, 7 (2010).

129. Bonis, P.A. *et al.* Hereditary nonpolyposis colorectal cancer: diagnostic strategies and their implications. *Evid Rep Technol Assess (Full Rep)*, 1-180 (2007).

130. Lynch, H.T., Shaw, M.W., Magnuson, C.W., Larsen, A.L. & Krush, A.J. Hereditary factors in cancer. Study of two large midwestern kindreds. *Arch Intern Med* **117**, 206-12 (1966).

131. Vasen, H.F., Mecklin, J.P., Khan, P.M. & Lynch, H.T. The International Collaborative Group on Hereditary Non-Polyposis Colorectal Cancer (ICG-HNPCC). *Dis Colon Rectum* **34**, 424-5 (1991).

132. Mecklin, J.P. & Jarvinen, H.J. Tumor spectrum in cancer family syndrome (hereditary nonpolyposis colorectal cancer). *Cancer* **68**, 1109-12 (1991).

133. Mitchell, R.J., Farrington, S.M., Dunlop, M.G. & Campbell, H. Mismatch repair genes hMLH1 and hMSH2 and colorectal cancer: a HuGE review. *Am J Epidemiol* **156**, 885-902 (2002).

134. Watson, P. & Lynch, H.T. Cancer risk in mismatch repair gene mutation carriers. *Fam Cancer* **1**, 57-60 (2001).

135. Dunlop, M.G. *et al.* Cancer risk associated with germline DNA mismatch repair gene mutations. *Hum Mol Genet* **6**, 105-10 (1997).

136. Hampel, H. *et al.* Screening for Lynch syndrome (hereditary nonpolyposis colorectal cancer) among endometrial cancer patients. *Cancer Res* **66**, 7810-7 (2006).

137. Vasen, H.F., Watson, P., Mecklin, J.P. & Lynch, H.T. New clinical criteria for hereditary nonpolyposis colorectal cancer (HNPCC, Lynch syndrome) proposed by the International Collaborative group on HNPCC. *Gastroenterology* **116**, 1453-6 (1999).

138. Rodriguez-Bigas, M.A. *et al.* A National Cancer Institute Workshop on Hereditary Nonpolyposis Colorectal Cancer Syndrome: meeting highlights and Bethesda guidelines. *J Natl Cancer Inst* **89**, 1758-62 (1997).

139. Umar, A. *et al.* Revised Bethesda Guidelines for hereditary nonpolyposis colorectal cancer (Lynch syndrome) and microsatellite instability. *J Natl Cancer Inst* **96**, 261-8 (2004).

140. Chen, P.C. *et al.* Contributions by MutL homologues Mlh3 and Pms2 to DNA mismatch repair and tumor suppression in the mouse. *Cancer Res* **65**, 8662-70 (2005).

141. Prolla, T.A. *et al.* Tumour susceptibility and spontaneous mutation in mice deficient in Mlh1, Pms1 and Pms2 DNA mismatch repair. *Nat Genet* **18**, 276-9 (1998).

142. Papadopoulos, N. & Lindblom, A. Molecular basis of HNPCC: mutations of MMR genes. *Hum Mutat* **10**, 89-99 (1997).

143. Thompson, E. *et al.* Hereditary non-polyposis colorectal cancer and the role of hPMS2 and hEXO1 mutations. *Clin Genet* **65**, 215-25 (2004).

144. Ligtenberg, M.J. *et al.* 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 (2009).

145. Jiricny, J. The multifaceted mismatch-repair system. *Nat Rev Mol Cell Biol* **7**, 335-46 (2006).

146. Kunkel, T.A. & Erie, D.A. DNA mismatch repair. *Annu Rev Biochem* **74**, 681-710 (2005).

147. Peltomaki, P. Deficient DNA mismatch repair: a common etiologic factor for colon cancer. *Hum Mol Genet* **10**, 735-40 (2001).

148. Hoeijmakers, J.H. Genome maintenance mechanisms for preventing cancer. *Nature* **411**, 366-74 (2001).

149. Berndt, S.I. *et al.* Mismatch repair polymorphisms and the risk of colorectal cancer. *Int J Cancer* **120**, 1548-54 (2007).

150. Drummond, J.T., Li, G.M., Longley, M.J. & Modrich, P. Isolation of an hMSH2-p160 heterodimer that restores DNA mismatch repair to tumor cells. *Science* **268**, 1909-12 (1995).

151. Palombo, F. *et al.* GTBP, a 160-kilodalton protein essential for mismatch-binding activity in human cells. *Science* **268**, 1912-4 (1995).

152. Palombo, F. *et al.* hMutSbeta, a heterodimer of hMSH2 and hMSH3, binds to insertion/deletion loops in DNA. *Curr Biol* **6**, 1181-4 (1996).

153. Li, G.M. & Modrich, P. Restoration of mismatch repair to nuclear extracts of H6 colorectal tumor cells by a heterodimer of human MutL homologs. *Proc Natl Acad Sci U S A* **92**, 1950-4 (1995).

154. Kadyrov, F.A., Dzantiev, L., Constantin, N. & Modrich, P. Endonucleolytic function of MutLalpha in human mismatch repair. *Cell* **126**, 297-308 (2006).

155. Raschle, M., Marra, G., Nystrom-Lahti, M., Schar, P. & Jiricny, J. Identification of hMutLbeta, a heterodimer of hMLH1 and hPMS1. *J Biol Chem* **274**, 32368-75 (1999).

156. Cannavo, E. *et al.* Expression of the MutL homologue hMLH3 in human cells and its role in DNA mismatch repair. *Cancer Res* **65**, 10759-66 (2005).

157. Pluciennik, A. *et al.* PCNA function in the activation and strand direction of MutL{alpha} endonuclease in mismatch repair. *Proc Natl Acad Sci U S A* (2010).

158. Bowers, J., Tran, P.T., Joshi, A., Liskay, R.M. & Alani, E. MSH-MLH complexes formed at a DNA mismatch are disrupted by the PCNA sliding clamp. *J Mol Biol* **306**, 957-68 (2001).

159. Clark, A.B., Valle, F., Drotschmann, K., Gary, R.K. & Kunkel, T.A. Functional interaction of proliferating cell nuclear antigen with MSH2-MSH6 and MSH2-MSH3 complexes. *J Biol Chem* **275**, 36498-501 (2000).

160. Gu, L., Hong, Y., McCulloch, S., Watanabe, H. & Li, G.M. ATP-dependent interaction of human mismatch repair proteins and dual role of PCNA in mismatch repair. *Nucleic Acids Res* **26**, 1173-8 (1998).

161. Kleczkowska, H.E., Marra, G., Lettieri, T. & Jiricny, J. hMSH3 and hMSH6 interact with PCNA and colocalize with it to replication foci. *Genes Dev* **15**, 724-36 (2001).

162. Umar, A. *et al.* Requirement for PCNA in DNA mismatch repair at a step preceding DNA resynthesis. *Cell* **87**, 65-73 (1996).

163. Gatti, R.A. The inherited basis of human radiosensitivity. *Acta Oncol* **40**, 702-11 (2001).

164. Popanda, O. *et al.* Radiation-induced DNA damage and repair in lymphocytes from breast cancer patients and their correlation with acute skin reactions to radiotherapy. *Int J Radiat Oncol Biol Phys* **55**, 1216-25 (2003).

165. Angele, S. *et al.* ATM haplotypes and cellular response to DNA damage: association with breast cancer risk and clinical radiosensitivity. *Cancer Res* **63**, 8717-25 (2003).

166. Hermanson-Miller, I.L. & Turchi, J.J. Strand-specific binding of RPA and XPA to damaged duplex DNA. *Biochemistry* **41**, 2402-8 (2002).

167. Iakoucheva, L.M., Walker, R.K., van Houten, B. & Ackerman, E.J. Equilibrium and stop-flow kinetic studies of fluorescently labeled DNA substrates with DNA repair proteins XPA and replication protein A. *Biochemistry* **41**, 131-43 (2002).

168. Iftode, C. & Borowiec, J.A. 5' --> 3' molecular polarity of human replication protein A (hRPA) binding to pseudo-origin DNA substrates. *Biochemistry* **39**, 11970-81 (2000).

169. Lao, Y., Lee, C.G. & Wold, M.S. Replication protein A interactions with DNA. 2. Characterization of double-stranded DNA-binding/helix-destabilization activities and the role of the zinc-finger domain in DNA interactions. *Biochemistry* **38**, 3974-84 (1999).

170. Lavrik, O.I., Kolpashchikov, D.M., Nasheuer, H.P., Weisshart, K. & Favre, A. Alternative conformations of human replication protein A are detected by crosslinks with primers carrying a photoreactive group at the 3'-end. *FEBS Lett* **441**, 186-90 (1998).

171. Treuner, K., Ramsperger, U. & Knippers, R. Replication protein A induces the unwinding of long double-stranded DNA regions. *J Mol Biol* **259**, 104-12 (1996).

172. Liu, L., Mo, J., Rodriguez-Belmonte, E.M. & Lee, M.Y. Identification of a fourth subunit of mammalian DNA polymerase delta. *J Biol Chem* **275**, 18739-44 (2000).

173. Podust, V.N., Chang, L.S., Ott, R., Dianov, G.L. & Fanning, E. Reconstitution of human DNA polymerase delta using recombinant baculoviruses: the p12 subunit potentiates DNA polymerizing activity of the four-subunit enzyme. *J Biol Chem* **277**, 3894-901 (2002).

174. Li, H. *et al.* Functional roles of p12, the fourth subunit of human DNA polymerase delta. *J Biol Chem* **281**, 14748-55 (2006).

175. Saribasak, H., Rajagopal, D., Maul, R.W. & Gearhart, P.J. Hijacked DNA repair proteins and unchained DNA polymerases. *Philos Trans R Soc Lond B Biol Sci* **364**, 605-11 (2009).

176. Chung, D.C. & Rustgi, A.K. The hereditary nonpolyposis colorectal cancer syndrome: genetics and clinical implications. *Ann Intern Med* **138**, 560-70 (2003).

177. Sutter, C., Gebert, J., Bischoff, P., Herfarth, C. & von Knebel Doeberitz, M. Molecular screening of potential HNPCC patients using a multiplex microsatellite PCR system. *Mol Cell Probes* **13**, 157-65 (1999).

178. Kolodner, R.D. *et al.* Germ-line msh6 mutations in colorectal cancer families. *Cancer Res* **59**, 5068-74 (1999).

179. Nystrom-Lahti, M. *et al.* DNA mismatch repair gene mutations in 55 kindreds with verified or putative hereditary non-polyposis colorectal cancer. *Hum Mol Genet* **5**, 763-9 (1996).

180. Park, J.G. *et al.* Suspected hereditary nonpolyposis colorectal cancer: International Collaborative Group on Hereditary Non-Polyposis Colorectal Cancer (ICG-HNPCC) criteria and results of genetic diagnosis. *Dis Colon Rectum* **42**, 710-5; discussion 715-6 (1999).

181. Thibodeau, S.N. *et al.* Altered expression of hMSH2 and hMLH1 in tumors with microsatellite instability and genetic alterations in mismatch repair genes. *Cancer Res* **56**, 4836-40 (1996).

182. Kemp, Z., Thirlwell, C., Sieber, O., Silver, A. & Tomlinson, I. An update on the genetics of colorectal cancer. *Hum Mol Genet* **13 Spec No 2**, R177-85 (2004).

183. McPhillips, M., Meldrum, C.J., Creegan, R., Edkins, E. & Scott, R.J. Deletion Mutations in an Australian Series of HNPCC Patients. *Hered Cancer Clin Pract* **3**, 43-7 (2005).

184. Macias, H. & Hinck, L. Mammary gland development. *Wiley Interdiscip Rev Dev Biol* **1**, 533-57 (2012).

185. Clendenen, T.V. *et al.* Magnetic resonance imaging (MRI) of hormone-induced breast changes in young premenopausal women. *Magn Reson Imaging* **31**, 1-9 (2013).

186. Sinn, H.P. & Kreipe, H. A Brief Overview of the WHO Classification of Breast Tumors, 4th Edition, Focusing on Issues and Updates from the 3rd Edition. *Breast Care (Basel)* **8**, 149-54 (2013).

187. Viale, G. The current state of breast cancer classification. *Ann Oncol* **23 Suppl 10**, x207-10 (2012).

188. Kroman, N., Wohlfahrt, J., Mouridsen, H.T. & Melbye, M. Influence of tumor location on breast cancer prognosis. *Int J Cancer* **105**, 542-5 (2003).

189. Dietel, M. Hormone replacement therapy (HRT), breast cancer and tumor pathology. *Maturitas* **65**, 183-9 (2010).

190. Stevenson, J.C. Hormone replacement therapy: review, update, and remaining questions after the Women's Health Initiative Study. *Curr Osteoporos Rep* **2**, 12-6 (2004).

191. Verkooijen, H.M., Bouchardy, C., Vinh-Hung, V., Rapiti, E. & Hartman, M. The incidence of breast cancer and changes in the use of hormone replacement therapy: a review of the evidence. *Maturitas* **64**, 80-5 (2009).

192. Glass, A.G., Lacey, J.V., Jr., Carreon, J.D. & Hoover, R.N. Breast cancer incidence, 1980-2006: combined roles of menopausal hormone therapy, screening mammography, and estrogen receptor status. *J Natl Cancer Inst* **99**, 1152-61 (2007).

193. Bernstein, L. & Ross, R.K. Endogenous hormones and breast cancer risk. *Epidemiol Rev* **15**, 48-65 (1993).

194. Key, T.J. & Pike, M.C. The role of oestrogens and progestagens in the epidemiology and prevention of breast cancer. *Eur J Cancer Clin Oncol* **24**, 29-43 (1988).

195. Anderson, E. The role of oestrogen and progesterone receptors in human mammary development and tumorigenesis. *Breast Cancer Res* **4**, 197-201 (2002).

196. Henderson, B.E., Ross, R.K., Pike, M.C. & Casagrande, J.T. Endogenous hormones as a major factor in human cancer. *Cancer Res* **42**, 3232-9 (1982).

197. Dunnwald, L.K., Rossing, M.A. & Li, C.I. Hormone receptor status, tumor characteristics, and prognosis: a prospective cohort of breast cancer patients. *Breast Cancer Res* **9**, R6 (2007).

198. Paul Wright, G., Davis, A.T., Koehler, T.J., Melnik, M.K. & Chung, M.H. Hormone receptor status does not affect prognosis in metaplastic breast cancer: a population-based analysis with comparison to infiltrating ductal and lobular carcinomas. *Ann Surg Oncol* **21**, 3497-503 (2014).

199. Giuliano, A.E. Mapping a pathway for axillary staging: a personal perspective on the current status of sentinel lymph node dissection for breast cancer. *Arch Surg* **134**, 195-9 (1999).

200. Grotenhuis, B.A., Klem, T.M. & Vrijland, W.W. Treatment outcome in breast cancer patients with ipsilateral supraclavicular lymph node metastasis at time of diagnosis: a review of the literature. *Eur J Surg Oncol* **39**, 207-12 (2013).

201. Newman, L.A. *et al.* Adverse prognostic significance of infraclavicular lymph nodes detected by ultrasonography in patients with locally advanced breast cancer. *Am J Surg* **181**, 313-8 (2001).

202. Vidal-Sicart, S. & Valdes Olmos, R. Sentinel node mapping for breast cancer: current situation. *J Oncol* **2012**, 361341 (2012).

203. Vrana, D., Gatek, J., Cwiertka, K., Lukesova, L. & Koranda, P. Internal mammary node management in breast cancer. A review. *Biomed Pap Med Fac Univ Palacky Olomouc Czech Repub* **157**, 261-5 (2013).

204. Scully, O.J., Bay, B.H., Yip, G. & Yu, Y. Breast cancer metastasis. *Cancer Genomics Proteomics* **9**, 311-20 (2012).

205. Stratton, M.R. & Rahman, N. The emerging landscape of breast cancer susceptibility. *Nat Genet* **40**, 17-22 (2008).

206. Mourouti, N., Kontogianni, M.D., Papavagelis, C. & Panagiotakos, D.B. Diet and breast cancer: a systematic review. *Int J Food Sci Nutr* **66**, 1-42 (2015).

207. Nickels, S. *et al.* Evidence of gene-environment interactions between common breast cancer susceptibility loci and established environmental risk factors. *PLoS Genet* **9**, e1003284 (2013).

208. Wu, Y.C., Zheng, D., Sun, J.J., Zou, Z.K. & Ma, Z.L. Meta-analysis of studies on breast cancer risk and diet in Chinese women. *Int J Clin Exp Med* **8**, 73-85 (2015).

209. Harvie, M., Howell, A. & Evans, D.G. Can diet and lifestyle prevent breast cancer: what is the evidence? *Am Soc Clin Oncol Educ Book* **35**, e66-73 (2015).

210. Khanna, K.K. & Jackson, S.P. DNA double-strand breaks: signaling, repair and the cancer connection. *Nat Genet* **27**, 247-54 (2001).

211. Lees-Miller, S.P. & Meek, K. Repair of DNA double strand breaks by non-homologous end joining. *Biochimie* **85**, 1161-73 (2003).

212. Harewood, L., Chaignat, E. & Reymond, A. Structural variation and its effect on expression. *Methods Mol Biol* **838**, 173-86 (2012).

213. Brandsma, I. & Gent, D.C. Pathway choice in DNA double strand break repair: observations of a balancing act. *Genome Integr* **3**, 9 (2012).

214. Grabarz, A., Barascu, A., Guirouilh-Barbat, J. & Lopez, B.S. Initiation of DNA double strand break repair: signaling and single-stranded resection dictate the choice between homologous recombination, non-homologous end-joining and alternative end-joining. *Am J Cancer Res* **2**, 249-68 (2012).

215. Pardo, B., Gomez-Gonzalez, B. & Aguilera, A. DNA repair in mammalian cells: DNA double-strand break repair: how to fix a broken relationship. *Cell Mol Life Sci* **66**, 1039-56 (2009).

216. Delacote, F. & Lopez, B.S. Importance of the cell cycle phase for the choice of the appropriate DSB repair pathway, for genome stability maintenance: the trans-S double-strand break repair model. *Cell Cycle* **7**, 33-8 (2008).

217. Lalloo, F. & Evans, D.G. Familial breast cancer. *Clin Genet* **82**, 105-14 (2012).

218. Peto, J. & Mack, T.M. High constant incidence in twins and other relatives of women with breast cancer. *Nat Genet* **26**, 411-4 (2000).

219. Easton, D.F. Familial risks of breast cancer. *Breast Cancer Res* **4**, 179-81 (2002).

220. Gracia-Aznarez, F.J. *et al.* Whole Exome Sequencing Suggests Much of Non-BRCA1/BRCA2 Familial Breast Cancer Is Due to Moderate and Low Penetrance Susceptibility Alleles. *PLoS One* **8**, e55681 (2013).

221. Wong, M.W. *et al.* BRIP1, PALB2, and RAD51C mutation analysis reveals their relative importance as genetic susceptibility factors for breast cancer. *Breast Cancer Res Treat* **127**, 853-9 (2011).

222. Antoniou, A. *et al.* Average risks of breast and ovarian cancer associated with BRCA1 or BRCA2 mutations detected in case Series unselected for family history: a combined analysis of 22 studies. *Am J Hum Genet* **72**, 1117-30 (2003).

223. Mavaddat, N., Antoniou, A.C., Easton, D.F. & Garcia-Closas, M. Genetic susceptibility to breast cancer. *Mol Oncol* **4**, 174-91 (2010).

224. Thompson, D. & Easton, D. The genetic epidemiology of breast cancer genes. *J Mammary Gland Biol Neoplasia* **9**, 221-36 (2004).

225. Zhang, B., Beeghly-Fadiel, A., Long, J. & Zheng, W. Genetic variants associated with breast-cancer risk: comprehensive research synopsis, meta-analysis, and epidemiological evidence. *Lancet Oncol* **12**, 477-88 (2011).

226. Michailidou, K. *et al.* Large-scale genotyping identifies 41 new loci associated with breast cancer risk. *Nat Genet* **45**, 353-61 (2013).

227. Feuk, L., Carson, A.R. & Scherer, S.W. Structural variation in the human genome. *Nat Rev Genet* **7**, 85-97 (2006).

228. Przeworski, M., Hudson, R.R. & Di Rienzo, A. Adjusting the focus on human variation. *Trends Genet* **16**, 296-302 (2000).

229. Reich, D.E. & Lander, E.S. On the allelic spectrum of human disease. *Trends Genet* **17**, 502-10 (2001).

230. Nei, M. Selectionism and Neutralism in Molecular Evolution. *Mol Biol Evol.* **22**, 24 (2005).

231. Kaneko, K. & Furusawa, C. An evolutionary relationship between genetic variation and phenotypic fluctuation. *J Theor Biol* **240**, 78-86 (2006).

232. Vida, G. Global issues of genetic diversity. *Exs* **68**, 9-19 (1994).

233. Rovelet-Lecrux, A. *et al.* APP locus duplication causes autosomal dominant early-onset Alzheimer disease with cerebral amyloid angiopathy. *Nat Genet* **38**, 24-6 (2006).

234. Singleton, A.B. *et al.* alpha-Synuclein locus triplication causes Parkinson's disease. *Science* **302**, 841 (2003).

235. Wang, Y. *et al.* Mutation in Rpa1 results in defective DNA double-strand break repair, chromosomal instability and cancer in mice. *Nat Genet* **37**, 750-5 (2005).

236. Fujii, K., Shikazono, N. & Yokoya, A. Nucleobase lesions and strand breaks in dry DNA thin film selectively induced by monochromatic soft X-rays. *J Phys Chem B* **113**, 16007-15 (2009).

237. Purschke, M., Laubach, H.J., Anderson, R.R. & Manstein, D. Thermal injury causes DNA damage and lethality in unheated surrounding cells: active thermal bystander effect. *J Invest Dermatol* **130**, 86-92 (2010).

238. Veierod, M.B., Adami, H.O., Lund, E., Armstrong, B.K. & Weiderpass, E. Sun and solarium exposure and melanoma risk: effects of age, pigmentary characteristics, and nevi. *Cancer Epidemiol Biomarkers Prev* **19**, 111-20 (2010).

239. Sedelnikova, O.A. *et al.* Role of oxidatively induced DNA lesions in human pathogenesis. *Mutat Res* **704**, 152-9 (2010).

240. Toyokuni, S. Molecular mechanisms of oxidative stress-induced carcinogenesis: from epidemiology to oxygenomics. *IUBMB Life* **60**, 441-7 (2008).

241. Hakem, R. DNA-damaged repair; the good, the bad and the ugly. *The EMBO Journal* **27**, 16 (2008).

242. Jackson, S.P. & Bartek, J. The DNA-damaged response in human biology and disease. *Nature* **461**, 7 (2009).

243. Paz-Elizur, T. *et al.* DNA repair of oxidative DNA damage in human carcinogenesis: potential application for cancer risk assessment and prevention. *Cancer Lett* **266**, 60-72 (2008).

244. Sutherland, B.M., Bennett, P.V., Sutherland, J.C. & Laval, J. Clustered DNA damages induced by x rays in human cells. *Radiat Res.* **157**, 5 (2002).

245. Axelsson, J., Bonde, J.P., Giwercman, Y.L., Rylander, L. & Giwercman, A. Gene-environment interaction and male reproductive function. *Asian J Androl* (2010).

246. Hung, R.J. *et al.* GST, NAT, SULT1A1, CYP1B1 genetic polymorphisms, interactions with environmental exposures and bladder cancer risk in a high-risk population. *Int J Cancer* **110**, 598-604 (2004).

247. Kim, J.I. *et al.* hOGG1 Ser326Cys polymorphism modifies the significance of the environmental risk factor for colon cancer. *World J Gastroenterol* **9**, 956-60 (2003).

248. Wynder, E.L. & Gori, G.B. Contribution of the environment to cancer incidence: an epidemiologic exercise. *J Natl Cancer Inst* **58**, 825-32 (1977).

249. Smith, P., McGuffog, L. & Easton, D.F. A genome wide linkage seach for breast cancer susceptability genes. *Genes Chromosomes Cancer* **45**, 9 (2006).

250. Win, A.K. *et al.* Determining the frequency of de novo germline mutations in DNA mismatch repair genes. *J Med Genet* **48**, 530-4 (2011).

251. Maiti, S., Kumar, K.H.B.G., Castellani, C.A., O'Reilly, R. & Singh, S.M. Ontogenetic De Novo Copy Number Variations (CNVs) as a Source of Genetic Individuality: Studies on Two Families with MZD Twins for Schizophrenia. *PLoS ONE* **6**(2011).

252. Sebat, J. *et al.* Strong Association of De Novo Copy Number Mutations with Autism. *Science* **316**, 4 (2007).

253. Bisgaard, M.L., Fenger, K., Bulow, S., Niebuhr, E. & Mohr, J. Familial adenomatous polyposis (FAP): frequency, penetrance, and mutation rate. *Hum Mutat* **3**, 121-5 (1994).

254. Ripa, R., Bisgaard, M.L., Bulow, S. & Nielsen, F.C. De novo mutations in familial adenomatous polyposis (FAP). *Eur J Hum Genet* **10**, 631-7 (2002).

255. Knudson, A.G., Jr. Mutation and cancer: statistical study of retinoblastoma. *Proc Natl Acad Sci U S A* **68**, 820-3 (1971).

256. Toss, A. *et al.* Hereditary Ovarian Cancer: Not Only 1 and 2 Genes. *Biomed Res Int* **2015**, 341723 (2015).

257. Mester, J. & Eng, C. Cowden syndrome: recognizing and managing a not-so-rare hereditary cancer syndrome. *J Surg Oncol* **111**, 125-30 (2015).

258. Mantere, T. *et al.* Finnish Fanconi anemia mutations and hereditary predisposition to breast and prostate cancer. *Clin Genet* **88**, 68-73 (2015).

259. Macaron, C., Leach, B.H. & Burke, C.A. Hereditary colorectal cancer syndromes and genetic testing. *J Surg Oncol* **111**, 103-11 (2015).

260. Ford, J.M. Hereditary Gastric Cancer: An Update at 15 Years. *JAMA Oncol* **1**, 16-8 (2015).

261. Economopoulou, P., Dimitriadis, G. & Psyrri, A. Beyond BRCA: new hereditary breast cancer susceptibility genes. *Cancer Treat Rev* **41**, 1-8 (2015).

262. Cancer Genome Atlas, N. Comprehensive molecular characterization of human colon and rectal cancer. *Nature* **487**, 330-7 (2012).

263. Cancer Genome Atlas, N. Comprehensive molecular portraits of human breast tumours. *Nature* **490**, 61-70 (2012).

264. Collins, F.S., Brooks, L.D. & Chakravarti, A. A DNA polymorphism discovery resource for research on human genetic variation. *Genome Res* **8**, 1229-31 (1998).

265. Sundaram, S.S., Bove, K.E., Lovell, M.A. & Sokol, R.J. Mechanisms of disease: Inborn errors of bile acid synthesis. *Nat Clin Pract Gastroenterol Hepatol* **5**, 456-68 (2008).

266. Nomizu, T. *et al.* Three cases of kindred with familial breast cancer in which carrier detection by BRCA gene testing was performed on family members. *Breast Cancer* **Online**, http://www.springerlink.com/content/t1xp25720q445605/ (2009).

267. Merchant, A. *et al.* Somatic mutations in SQSTM1 detected in affected tissues from patients with sporadic Paget's disease of bone. *J Bone Miner Res* **24**, 484-94 (2009).

268. Sharp, A.J. *et al.* Segmental duplications and copy-number variation in the human genome. *Am J Hum Genet* **77**, 78-88 (2005).

269. Tuzun, E. *et al.* Fine-scale structural variation of the human genome. *Nat Genet* **37**, 727-32 (2005).

270. Sebat, J. *et al.* Large-scale copy number polymorphism in the human genome. *Science* **305**, 525-8 (2004).

271. International consortium completes human genome project. *Pharmacogenomics* **4**, 241 (2003).

272. Spitz, M.R. & Bondy, M.L. The evolving discipline of molecular epidemiology of cancer. *Carcinogenesis* **31**, 127-34 (2010).

273. Redon, R. *et al.* Global variation in copy number in the human genome. *Nature* **444**, 444-54 (2006).

274. Conrad, D.F., Andrews, T.D., Carter, N.P., Hurles, M.E. & Pritchard, J.K. A high-resolution survey of deletion polymorphism in the human genome. *Nat Genet* **38**, 75-81 (2006).

275. Hinds, D.A., Kloek, A.P., Jen, M., Chen, X. & Frazer, K.A. Common deletions and SNPs are in linkage disequilibrium in the human genome. *Nat Genet* **38**, 82-5 (2006).

276. McCarroll, S.A. *et al.* Common deletion polymorphisms in the human genome. *Nat Genet* **38**, 86-92 (2006).

277. Repping, S. *et al.* High mutation rates have driven extensive structural polymorphism among human Y chromosomes. *Nat Genet* **38**, 463-7 (2006).

278. Ionita-Laza, I., Rogers, A.J., Lange, C., Raby, B.A. & Lee, C. Genetic association analysis of copy-number variation (CNV) in human disease pathogenesis. *Genomics* **93**, 22-6 (2009).

279. de Vries, B.B. *et al.* Diagnostic genome profiling in mental retardation. *Am J Hum Genet* **77**, 606-16 (2005).

280. Schoumans, J. *et al.* Detection of chromosomal imbalances in children with idiopathic mental retardation by array based comparative genomic hybridisation (array-CGH). *J Med Genet* **42**, 699-705 (2005).

281. Tyson, C. *et al.* Submicroscopic deletions and duplications in individuals with intellectual disability detected by array-CGH. *Am J Med Genet A* **139**, 173-85 (2005).

282. Girirajan, S. & Eichler, E.E. Phenotypic variability and genetic susceptibility to genomic disorders. *Hum Mol Genet* **19**, R176-87 (2010).

283. Delnatte, C. *et al.* Contiguous gene deletion within chromosome arm 10q is associated with juvenile polyposis of infancy, reflecting cooperation between the BMPR1A and PTEN tumor-suppressor genes. *Am J Hum Genet* **78**, 1066-74 (2006).

284. Alonso-Espinaco, V. *et al.* Novel MLH1 duplication identified in Colombian families with Lynch syndrome. *Genet Med* **13**, 155-60 (2011).

285. Krepischi, A.C. *et al.* Germline DNA copy number variation in familial and early-onset breast cancer. *Breast Cancer Res* **14**, R24 (2012).

286. Chan, T.L. *et al.* A novel germline 1.8-kb deletion of hMLH1 mimicking alternative splicing: a founder mutation in the Chinese population. *Oncogene* **20**, 2976-81 (2001).

287. Nystrom-Lahti, M. *et al.* Founding mutations and Alu-mediated recombination in hereditary colon cancer. *Nat Med* **1**, 1203-6 (1995).

288. Stella, A. *et al.* Germline novel MSH2 deletions and a founder MSH2 deletion associated with anticipation effects in HNPCC. *Clin Genet* **71**, 130-9 (2007).

289. Plaschke, J., Ruschoff, J. & Schackert, H.K. Genomic rearrangements of hMSH6 contribute to the genetic predisposition in suspected hereditary non-polyposis colorectal cancer syndrome. *J Med Genet* **40**, 597-600 (2003).

290. Suehiro, Y. *et al.* Germline copy number variations associated with breast cancer susceptibility in a Japanese population. *Tumour Biol* **34**, 947-52 (2013).

291. Kuusisto, K.M. *et al.* copy number variation analysis in familial BRCA1/2-negative Finnish breast and ovarian cancer. *PLoS One* **8**, e71802 (2013).

292. Nielsen, K.V. *et al.* The value of TOP2A gene copy number variation as a biomarker in breast cancer: Update of DBCG trial 89D. *Acta Oncol* **47**, 725-34 (2008).

293. Tchatchou, S. & Burwinkel, B. Chromosome copy number variation and breast cancer risk. *Cytogenet Genome Res* **123**, 183-7 (2008).

294. Tommasi, S. *et al.* Gene copy number variation in male breast cancer by aCGH. *Cell Oncol (Dordr)* **34**, 467-73 (2011).

295. Morak, M. *et al.* Biallelic MLH1 SNP cDNA expression or constitutional promoter methylation can hide genomic rearrangements causing Lynch syndrome. *J Med Genet* **48**, 513-519 (2011).

296. Clendenning, M. *et al.* Mutation deep within an intron of MSH2 causes Lynch syndrome. *Fam Cancer* **10**, 297-301 (2011).

297. Giarola, M. *et al.* Screening for mutations of the APC gene in 66 Italian familial adenomatous polyposis patients: evidence for phenotypic differences in cases with and without identified mutation. *Hum Mutat* **13**, 116-23 (1999).

298. Veitia, R.A., Bottani, S. & Birchler, J.A. Gene dosage effects: nonlinearities, genetic interactions, and dosage compensation. *Trends Genet* **29**, 385-93 (2013).

299. Migicovsky, Z. & Kovalchuk, I. Epigenetic memory in mammals. *Frontiers in Genetics* **2**, 7 (2011).

300. Bonasio, R., Tu, S. & Reinberg, D. Molecular signals of epigenetic states. *Science* **330**, 4 (2010).

301. Wong, J.J., Hawkins, N.J. & Ward, R.L. CRC: A model for epigenetic tumour genesis. *Gut* **56**, 8 (2007).

302. Venkatachalam, R. *et al.* The epigenetics of (hereditary) colorectal cancer. *Cancer Genet Cytogenet* **203**, 1-6 (2011).

303. Chen, J. & Xu, X. Diet, epigenetic, and cancer prevention. *Adv Genet* **71**, 237-55 (2010).

304. Hardy, T.M. & Tollefsbol, T.O. Epigenetic diet: impact on the epigenome and cancer. *Epigenomics* **3**, 503-18 (2011).

305. Saleem, M. *et al.* Review-Epigenetic therapy for cancer. *Pak J Pharm Sci* **28**, 1023-32 (2015).

306. Vaish, V., Khare, T., Verma, M. & Khare, S. Epigenetic therapy for colorectal cancer. *Methods Mol Biol* **1238**, 771-82 (2015).

307. Auclair, J. *et al.* Intensity-dependent constitutional MLH1 promoter methylation leads to early onset of colorectal cancer by affecting both alleles. *Genes Chromosomes Cancer* **50**, 178-85 (2011).

308. Peltomaki, P. Mutations and epimutations in the origin of cancer. *Experimental Cell Research* **318**, 11 (2011).

309. Birgisdottir, V. *et al.* Epigenetic silencing and deletion of the BRCA1 gene in sporadic breast cancer. *Breast Cancer Res* **8**, R38 (2006).

310. Shukla, G.C., Singh, J. & Barik, S. MicroRNAs: Processing, Maturation, Target Recognition and Regulatory Functions. *Mol Cell Pharmacol* **3**, 9 (2011).

311. Filipowicz, W., Bhattacharyya, S.N. & Sonenberg, N. Mechanisms of post-transcriptional regulation by microRNAs: are the answers in sight? *Nat Rev Genet* **9**, 102-14 (2008).

312. Dumont, N. & Tlsty, T.D. Reflections on miR-ing effects in metastasis. *Cancer Cell* **16**, 3-4 (2009).

313. Shi, M., Liu, D., Duan, H., Shen, B. & Guo, N. Metastasis-related miRNAs, active players in breast cancer invasion, and metastasis. *Cancer Metastasis Rev* (2010).

314. Garofalo, M. & Croce, C.M. MicroRNAs: master regulators as potential therapeutics in cancer. *Annu Rev Pharmacol Toxicol* **51**, 18 (2011).

315. Lu, J. *et al.* MicroRNA expression profiles classify human cancers. *Nature* **435**, 834-8 (2005).

316. Rosenfeld, N. *et al.* MicroRNAs accurately identify cancer tissue origin. *Nat Biotechnol* **26**, 462-9 (2008).

317. Volinia, S. *et al.* A microRNA expression signature of human solid tumors defines cancer gene targets. *Proc Natl Acad Sci U S A* **103**, 2257-61 (2006).

318. Iorio, M.V. *et al.* MicroRNA gene expression deregulation in human breast cancer. *Cancer Res* **65**, 7065-70 (2005).

319. Nagel, R. *et al.* Regulation of the adenomatous polyposis coli gene by the miR-135 family in colorectal cancer. *Cancer Res* **68**, 5795-802 (2008).

320. Wang, X., Lam, E.K., Zhang, J., Jin, H. & Sung, J.J. MicroRNA-122a functions as a novel tumor suppressor downstream of adenomatous polyposis coli in gastrointestinal cancers. *Biochem Biophys Res Commun* **387**, 376-80 (2009).

321. Tang, J. *et al.* MicroRNA 345, a methylation-sensitive microRNA is involved in cell proliferation and invasion in human colorectal cancer. *Carcinogenesis* **Online**, 9 (2011).

322. Oberg, A.L. *et al.* miRNA Expression in Colon Polyps Provides Evidence for a Multihit Model of Colon Cancer. *PLoS One* **6**, e20465 (2010).

323. Motoyama, K. *et al.* Over- and under-expressed microRNAs in human colorectal cancer. *Int J Oncol* **34**, 1069-75 (2009).

324. Landi, D. *et al.* Polymorphisms within micro-RNA-binding sites and risk of sporadic colorectal cancer. *Carcinogenesis* **29**, 579-84 (2008).

325. Serpico, D., Molino, L. & Di Cosimo, S. microRNAs in breast cancer development and treatment. *Cancer Treat Rev* **40**, 595-604 (2014).

326. Bandiera, S., Hatem, E., Lyonnet, S. & Hanrion-Caude, A. MicroRNAs in disease from candiate to modifier genes. *Clinical Genetics* **77**, 7 (2010).

327. Zhang, L. *et al.* microRNAs exhibit high frequency genomic alterations in human cancer. *Proc Natl Acad Sci U S A* **103**, 9136-41 (2006).

328. Gilbert, W. Why genes in pieces? *Nature* **271**, 1 (1978).

329. Roy, S.W. & Gilbert, W. The evolution of spliceosomal introns: Patterns, puzzles and progress. *Nat Rev Genet* **7**, 10 (2006).

330. Uphoff, C.C., Habigs, S. & al, e. ABL-BCR expression and BCR-ABL positive human leukemia cell lines. *Leukemia Research* **23**, 5 (1999).

331. Liu, H. *et al.* Elucidating Translocation Gene Fusion by SNP Array Data. *Cancer Informatics* **11**, 12 (2012).

332. Tarrío, R., Ayala, F.J. & Rodríguez-Trelles, R. Alternative splicing: A missing piece in the puzzle

of intron gain. *PNAS* **105**, 5 (2008).

333. DeNoto, F.M., Moore, D.D. & Goodman, H.M. Human growth hormone DNA sequence and mRNA structure: Possible alternative splicing. *Nucleic Acids Res* **9**, 3719 (1981).

334. Early, P. & al., e. Two mRNAs can be produced from a single immunoglobulin mu gene by alternative RNA processing pathways. *Cell* **20**, 6 (1980).

335. Tuohy, T.M.F. *et al.* Large intron 14 rearrangement in APC results in splice defect and attenuated FAP. *Human Genetics* **127**, 10 (2010).

336. Mihalatos, M. *et al.* Rare mutations predisposing to familial adenomatous polyposis in Greek FAP patients. *BMC Cancer* **5**, 8 (2005).

337. Bianchi, F. *et al.* An intronic mutation in MLH1 associated with familial colon and breast cancer. *Familial Cancer* **10**, 8 (2011).

338. Liu, X., Sinn, H.P., Ulmer, H.U., Scott, R.J. & Hamann, U. Intronic TP53 Germline Sequence Variants Modify the Risk in German Breast/Ovarian Cancer Families. *Hered Cancer Clin Pract* **2**, 139-45 (2004).

339. Ratanaphan, A., Panomwan, P., Canyuk, B. & Maipang, T. Identification of novel intronic BRCA1 variants of uncertain significance in a Thai hereditary breast cancer family. *J Genet* **90**, 327-31 (2011).

340. Sun, T. *et al.* A six-nucleotide insertion-deletion polymorphism in the CASP8 promoter is associated with susceptibility to multiple cancers. *Nat Genet* **39**, 605-13 (2007).

341. Fokkema, I.F. *et al.* LOVD v.2.0: the next generation in gene variant databases. *Hum Mutat* **32**, 557-63 (2011).

342. Hochstenbach, R. *et al.* Discovery of variants unmasked by hemizygous deletions. *Eur J Hum Genet* **20**, 748-53 (2012).

343. Almal, S.H. & Padh, H. Implications of gene copy-number variation in health and diseases. *J Hum Genet* **57**, 6-13 (2012).

344. Bronstad, I., Wolff, A.S., Lovas, K., Knappskog, P.M. & Husebye, E.S. Genome-wide copy number variation (CNV) in patients with autoimmune Addison's disease. *BMC Med Genet* **12**, 111 (2011).

345. Grozeva, D. *et al.* Rare copy number variants: a point of rarity in genetic risk for bipolar disorder and schizophrenia. *Arch Gen Psychiatry* **67**, 318-27 (2010).

346. Hai, R. *et al.* Genome-wide association study of copy number variation identified gremlin1 as a candidate gene for lean body mass. *J Hum Genet* **57**, 33-7 (2012).

347. Jiang, Q., Ho, Y.Y., Hao, L., Nichols Berrios, C. & Chakravarti, A. Copy number variants in candidate genes are genetic modifiers of Hirschsprung disease. *PLoS One* **6**, e21219 (2011).

348. Wellcome Trust Case Control, C. *et al.* Genome-wide association study of CNVs in 16,000 cases of eight common diseases and 3,000 shared controls. *Nature* **464**, 713-20 (2010).

349. Nagasaka, T. *et al.* Somatic hypermethylation of MSH2 is a frequent event in Lynch Syndrome colorectal cancers. *Cancer Res* **70**, 3098-108 (2010).

350. Spier, I. *et al.* Deep intronic APC mutations explain a substantial proportion of patients with familial or early-onset adenomatous polyposis. *Hum Mutat* **33**, 1045-50 (2012).

351. Aretz, S. *et al.* Familial adenomatous polyposis: aberrant splicing due to missense or silent mutations in the APC gene. *Hum Mutat* **24**, 370-80 (2004).

352. Kaufmann, A. *et al.* Analysis of rare APC variants at the mRNA level: six pathogenic mutations and literature review. *J Mol Diagn* **11**, 131-9 (2009).

353. Auclair, J. *et al.* Systematic mRNA analysis for the effect of MLH1 and MSH2 missense and silent mutations on aberrant splicing. *Hum Mutat* **27**, 145-54 (2006).

354. Nathanson, K.L., Wooster, R. & Weber, B.L. Breast cancer genetics: what we know and what we need. *Nat Med* **7**, 552-6 (2001).

355. Wooster, R. & Weber, B.L. Breast and ovarian cancer. *N Engl J Med* **348**, 2339-47 (2003).

356. Lasko, D., Cavenee, W. & Nordenskjold, M. Loss of constitutional heterozygosity in human cancer. *Annu Rev Genet* **25**, 281-314 (1991).

357. Conrad, D.F. *et al.* Origins and functional impact of copy number variation in the human genome. *Nature* **464**, 704-12 (2010).

358. Shlien, A. & Malkin, D. Copy number variations and cancer susceptibility. *Curr Opin Oncol* **22**, 55-63 (2010).

359. Lu, M., Shi, B., Wang, J., Cao, Q. & Cui, Q. TAM: a method for enrichment and depletion analysis of a microRNA category in a list of microRNAs. *BMC Bioinformatics* **11**, 419 (2010).

360. Wang, J., Duncan, D., Shi, Z. & Zhang, B. WEB-based GEne SeT AnaLysis Toolkit (WebGestalt): update 2013. *Nucleic Acids Res* **41**, W77-83 (2013).

361. Zhang, B., Kirov, S. & Snoddy, J. WebGestalt: an integrated system for exploring gene sets in various biological contexts. *Nucleic Acids Res* **33**, W741-8 (2005).

362. Su, Y.H. *et al.* MIR-142-5p and miR-9 may be involved in squamous lung cancer by regulating cell cycle related genes. *Eur Rev Med Pharmacol Sci* **17**, 3213-20 (2013).

363. Jemal, A. *et al.* Global cancer statistics. *CA Cancer J Clin* **61**, 69-90 (2011).

364. Hanahan, D. & Weinberg, R.A. The hallmarks of cancer. *Cell* **100**, 57-70 (2000).

365. Dellinger, A.E. *et al.* Comparative analyses of seven algorithms for copy number variant identification from single nucleotide polymorphism arrays. *Nucleic Acids Res* **38**, e105 (2010).

366. Tsuang, D.W. *et al.* The effect of algorithms on copy number variant detection. *PLoS One* **5**, e14456 (2010).

367. Zhang, D. *et al.* Accuracy of CNV Detection from GWAS Data. *PLoS One* **6**, e14511 (2011).

368. Pinto, D. *et al.* Comprehensive assessment of array-based platforms and calling algorithms for detection of copy number variants. *Nat Biotechnol* **29**, 512-20 (2011).

369. Kim, S.Y., Kim, J.H. & Chung, Y.J. Effect of Combining Multiple CNV Defining Algorithms on the Reliability of CNV Calls from SNP Genotyping Data. *Genomics Inform* **10**, 194-9 (2012).

370. Murata, H., Khattar, N.H., Gu, L. & Li, G.M. Roles of mismatch repair proteins hMSH2 and hMLH1 in the development of sporadic breast cancer. *Cancer Lett* **223**, 143-50 (2005).

371. Vodusek, A.L., Novakovic, S., Stegel, V. & Jereb, B. Genotyping of BRCA1, BRCA2, p53, CDKN2A, MLH1 and MSH2 genes in a male patient with secondary breast cancer. *Radiol Oncol* **45**, 296-9 (2011).

372. Bartkova, J. *et al.* Aberrations of the MRE11-RAD50-NBS1 DNA damage sensor complex in human breast cancer: MRE11 as a candidate familial cancer-predisposing gene. *Mol Oncol* **2**, 296-316 (2008).

373. Heikkinen, K., Karppinen, S.M., Soini, Y., Makinen, M. & Winqvist, R. Mutation screening of Mre11 complex genes: indication of RAD50 involvement in breast and ovarian cancer susceptibility. *J Med Genet* **40**, e131 (2003).

374. Hsu, H.M. *et al.* Breast cancer risk is associated with the genes encoding the DNA double-strand break repair Mre11/Rad50/Nbs1 complex. *Cancer Epidemiol Biomarkers Prev* **16**, 2024-32 (2007).

375. Yuan, S.S. *et al.* Role of MRE11 in cell proliferation, tumor invasion, and DNA repair in breast cancer. *J Natl Cancer Inst* **104**, 1485-502 (2012).

376. Gothlin Eremo, A. *et al.* Wwox expression may predict benefit from adjuvant tamoxifen in randomized breast cancer patients. *Oncol Rep* **29**, 1467-74 (2013).

377. Ekizoglu, S., Muslumanoglu, M., Dalay, N. & Buyru, N. Genetic alterations of the WWOX gene in breast cancer. *Med Oncol* **29**, 1529-35 (2012).

378. Lucito, R. *et al.* Copy-number variants in patients with a strong family history of pancreatic cancer. *Cancer Biol Ther* **6**, 1592-9 (2007).

379. Campiglio, M. *et al.* FHIT loss of function in human primary breast cancer correlates with advanced stage of the disease. *Cancer Res* **59**, 3866-9 (1999).

380. Cecener, G. *et al.* Importance of novel sequence alterations in the FHIT gene on formation of breast cancer. *Tumori* **93**, 597-603 (2007).

381. Iliopoulos, D. *et al.* Roles of FHIT and WWOX fragile genes in cancer. *Cancer Lett* **232**, 27-36 (2006).

382. Ismail, H.M., Medhat, A.M., Karim, A.M. & Zakhary, N.I. Multiple Patterns of FHIT Gene Homozygous Deletion in Egyptian Breast Cancer Patients. *Int J Breast Cancer* **2011**, 325947 (2011).

383. Ismail, H.M., Medhat, A.M., Karim, A.M. & Zakhary, N.I. FHIT gene and flanking region on chromosome 3p are subjected to extensive allelic loss in Egyptian breast cancer patients. *Mol Carcinog* **50**, 625-34 (2011).

384. Negrini, M. *et al.* The FHIT gene at 3p14.2 is abnormal in breast carcinomas. *Cancer Res* **56**, 3173-9 (1996).

385. Bianchi, F., Tagliabue, E., Menard, S. & Campiglio, M. Fhit expression protects against HER2-driven breast tumor development: unraveling the molecular interconnections. *Cell Cycle* **6**, 643-6 (2007).

386. Arun, B. *et al.* Loss of FHIT expression in breast cancer is correlated with poor prognostic markers. *Cancer Epidemiol Biomarkers Prev* **14**, 1681-5 (2005).

387. Talseth-Palmer, B.A. *et al.* Continuing difficulties in interpreting CNV data: lessons from a genome-wide CNV association study of Australian HNPCC/lynch syndrome patients. *BMC Med Genomics* **6**, 10 (2013).

388. DeBerardinis, R.J. & Thompson, C.B. Cellular metabolism and disease: what do metabolic outliers teach us? *Cell* **148**, 1132-44 (2012).

389. Munoz-Pinedo, C., El Mjiyad, N. & Ricci, J.E. Cancer metabolism: current perspectives and future directions. *Cell Death Dis* **3**, e248 (2012).

390. Anczukow, O. *et al.* BRCA2 deep intronic mutation causing activation of a cryptic exon: opening toward a new preventive therapeutic strategy. *Clin Cancer Res* **18**, 4903-9 (2012).

391. De Moor, M.H. *et al.* Genome-wide association study of exercise behavior in Dutch and American adults. *Med Sci Sports Exerc* **41**, 1887-95 (2009).

392. Ferreira, M.A. *et al.* Collaborative genome-wide association analysis supports a role for ANK3 and CACNA1C in bipolar disorder. *Nat Genet* **40**, 1056-8 (2008).

393. Trevino, L.R. *et al.* Germline genomic variants associated with childhood acute lymphoblastic leukemia. *Nat Genet* **41**, 1001-5 (2009).

394. Costi, R. *et al.* Repeated anastomotic recurrence of colorectal tumors: genetic analysis of two cases. *World J Gastroenterol* **17**, 3752-8 (2011).

395. Shi, Z.Z. *et al.* Genomic profiling of rectal adenoma and carcinoma by array-based comparative genomic hybridization. *BMC Med Genomics* **5**, 52 (2012).

396. Loewer, S. *et al.* Large intergenic non-coding RNA-RoR modulates reprogramming of human induced pluripotent stem cells. *Nat Genet* **42**, 1113-7 (2010).

397. Nagano, T. *et al.* The Air noncoding RNA epigenetically silences transcription by targeting G9a to chromatin. *Science* **322**, 1717-20 (2008).

398. Orom, U.A. *et al.* Long noncoding RNAs with enhancer-like function in human cells. *Cell* **143**, 46-58 (2010).

399. Ye, L.C., Zhu, X., Qiu, J.J., Xu, J. & Wei, Y. Involvement of long non-coding RNA in colorectal cancer: From benchtop to bedside (Review). *Oncol Lett* **9**, 1039-1045 (2015).

400. Ma, Y. *et al.* Long non-coding RNA CCAL regulates colorectal cancer progression by activating Wnt/beta-catenin signalling pathway via suppression of activator protein 2alpha. *Gut* (2015).

401. Buerki, N. *et al.* Evidence for breast cancer as an integral part of Lynch syndrome. *Genes Chromosomes Cancer* **51**, 83-91 (2012).

402. Dowty, J.G. *et al.* Cancer risks for MLH1 and MSH2 mutation carriers. *Hum Mutat* **34**, 490-7 (2013).

403. Vasen, H.F., Morreau, H. & Nortier, J.W. Is breast cancer part of the tumor spectrum of hereditary nonpolyposis colorectal cancer? *Am J Hum Genet* **68**, 1533-5 (2001).

404. Shanley, S. *et al.* Breast cancer immunohistochemistry can be useful in triage of some HNPCC families. *Fam Cancer* **8**, 251-5 (2009).

405. Bajrami, I. *et al.* Synthetic lethality of PARP and NAMPT inhibition in triple-negative breast cancer cells. *EMBO Mol Med* **4**, 1087-96 (2012).

406. Lv, X. *et al.* Regulative Effect of Nampt on Tumor Progression and Cell Viability in Human Colorectal Cancer. *J Cancer* **6**, 849-58 (2015).

407. Neubauer, K. *et al.* Nampt/PBEF/visfatin upregulation in colorectal tumors, mirrored in normal tissue and whole blood of colorectal cancer patients, is associated with metastasis, hypoxia, IL1beta, and anemia. *Biomed Res Int* **2015**, 523930 (2015).

408. Zhang, C., Tong, J. & Huang, G. Nicotinamide phosphoribosyl transferase (Nampt) is a target of microRNA-26b in colorectal cancer cells. *PLoS One* **8**, e69963 (2013).

409. Blum, C. *et al.* The expression ratio of Map7/B2M is prognostic for survival in patients with stage II colon cancer. *Int J Oncol* **33**, 579-84 (2008).