View metadata, citation and similar papers at core.ac.uk
brought to you by CORE
provided by Helsingin yliopiston digitaalinen arkisto

Helsinki University Biomedical Dissertations No. 195

New insights into the genetic basis of colorectal cancer

Alexandra Gylfe, M.Sc.

Haartman Institute,
Department of Medical Genetics
Research Programs Unit,
Genome-Scale Biology Research Program

Faculty of Medicine
Helsinki Biomedical Graduate Program

University of Helsinki
Finland

Academic dissertation

To be publicly discussed with the permission of the Faculty of Medicine of the University of Helsinki, in Haartman Institute, Small Lecture Hall, Haartmaninkatu 8, Helsinki, on the 23rd of May 2014, at 12 noon.

Helsinki 2014

Thesis supervisors

Academy Professor Lauri A. Aaltonen, M.D., Ph.D.
Haartman Institute, Deparment of Medical Genetics
Research Programs Unit, Genome-Scale Biology Research Program
Faculty of Medicine
University of Helsinki
Finland

and

Docent Auli Karhu, Ph.D.
Haartman Institute, Deparment of Medical Genetics
Research Programs Unit, Genome-Scale Biology Research Program
Faculty of Medicine
University of Helsinki
Finland

Reviewers appointed by the Faculty

Docent Iirs Hovatta, Ph.D.
Department of Biosciences
University of Helsinki
Finland

and

Docent Pipsa Saharinen, Ph.D.
Research Programs Unit,
Translational Cancer Biology Research Program
University of Helsinki
Finland

Opponent appointed by the Faculty

Professor Nickolas Papadopoulos, Ph.D.
Department of Oncology
Ludwig Center for Cancer Genetics & Therapeutics
Sidney Kimmel Comprehensive Cancer Center
The Johns Hopkins University School of Medicine
United States

ISBN 978-952-10-9838-3 (paperback)
ISBN 978-952-10-9839-0 (PDF)
ISSN 1457-8433
http://ethesis.helsinki.fi
Unigrafia Oy
Helsinki 2014

To my family

TABLE OF CONTENTS

LIST OF ORIGINAL PUBLICATIONS

This thesis is based on the following publications that are referred to in the text by their Roman numerals.

I. Niittymäki I, **Gylfe A**, Laine L, Laakso M, Lehtonen HJ, Kondelin J, Tolvanen J, Nousiainen K, Pouwels J, Järvinen H, Nuorva K, Mecklin JP, Mäkinen M, Ristimäki A, Ørntoft TF, Hautaniemi S, Karhu A, Kallio MJ, Aaltonen LA. High frequency of TTK mutations in microsatellite-unstable colorectal cancer and evaluation of their effect on spindle assembly checkpoint. *Carcinogenesis.* 2011 32:305-11.

II. **Gylfe AE***, Kondelin J*, Turunen M, Ristolainen H, Katainen R, Pitkänen E, Kaasinen E, Rantanen V, Tanskanen T, Varjosalo M, Lehtonen H, Palin K, Taipale M, Taipale J, Renkonen-Sinisalo L, Järvinen H, Böhm J, Mecklin JP, Ristimäki A, Kilpivaara O, Tuupanen S, Karhu A, Vahteristo P, Aaltonen LA. Identification of Candidate Oncogenes in Human Colorectal Cancers with Microsatellite Instability. *Gastroenterology.* 2013 145:540-543.

III. **Gylfe AE**, Sirkiä J, Ahlsten M, Järvinen H, Mecklin JP, Karhu A, Aaltonen LA. Somatic mutations and germline sequence variants in patients with familial colorectal cancer. *Int J Cancer*. 2010 127:2974-80.

IV. **Gylfe AE,** Katainen R, Kondelin J, Tanskanen T, Cajuso T, Hänninen U, Taipale J, Taipale M, Renkonen-Sinisalo L, Järvinen H, Mecklin JP, Kilpivaara P, Pitkänen E, Vahteristo P, Tuupanen S, Karhu A, Aaltonen LA. Eleven candidate susceptibility genes for common familial colorectal cancer. *Plos Genetics.* 2013 9(10):e1003876.

* Equal contribution

The publications are reproduced with the permission of the copyright holders. Publication I was included in the thesis of Iina Niittymäki (Molecular features of colorectal cancer - predisposition and progression, Helsinki 2011).

ABBREVIATIONS

ACF	aberrant crypt foci
APC	adenomatous polyposis coli protein
ASPP2	apoptosis-stimulating of p53 protein 2
BAX	apoptosis regulator BAX
BLM	bloom syndrome protein
BRAF	serine/threonine-protein kinase B-raf
BUB1	mitotic checkpoint serine/threonine protein kinase BUB1
CAN	candidate cancer genes
cDNA	complementary deoxyribonucleic acid
CIN	chromosomal instability
CK1	casein kinase I isoform alpha
CRC	colorectal cancer
DMSO	dimethyl sulfoxide
DNA	deoxyribonucleic acid
EDM	exonuclease domain mutations
EDTA	ethylene-diamine-tetraacetic acid
EGF	epidermal growth factor
EGFR	epidermal growth factor receptor
ERK	extracellular signal-regulated kinase
FAP	familial adenomatous polyposis
FBXW7	f-box/WD repeat-containing protein 7
GFP	green fluorescent protein
GSK-3β	glycogen synthase kinase-3 beta
GTP	guanosine-5'-triphosphate
GWA	genome-wide association
HA	hemagglutinin
HCL	hydrochloric acid
HDM2	E3 ubiquitin-protein ligase Mdm2
HNPCC	hereditary nonpolyposis colorectal cancer
JP	juvenile polyposis
KRAS	GTPase KRas
LEF	lymphoid enhancer factor
LoF	loss of functions
LOH	loss of heterozygosity
MAD2	mitotic spindle assembly checkpoint protein MAD2A
MAF	minor allele frequency
MAPK	mitogen-activated protein kinase
MEK	mitogen-activated protein kinase
MLH1	MutL protein homolog 1 (E. coli)
MLH3	MutL protein homolog 3
MMR	mismatch repair
mRNA	messenger ribonucleic acid
MSH2, 3, 6	mutS protein homolog 2, 3, 6 (E. coli)
MSI	microsatellite instability
MSS	microsatellite stable
NGS	next-generation sequencing
NMD	nonsense-mediated decay
PI3K	phosphatidylinositide 3-kinase
PJS	Peutz–Jeghers syndrome
PMS1	PMS1 protein homolog 1
PMS2	PMS1 protein homolog 2
POLD1	DNA polymerase delta catalytic subunit
POLE	DNA polymerase epsilon catalytic subunit A
PSRC1	proline/serine-rich coiled-coil protein 1
PTC	premature termination codon
RAF	RAF proto-oncogene serine/threonine protein kinase
RAN	GTP-binding nuclear protein Ran
RANBP2	E3 SUMO-protein ligase RanBP2
RB1	retinoblastoma-associated protein
SAC	spindle assembly checkpoint
SMAD	mothers against decapentaplegic homolog
TCF	t-cell factor
TCGA	the cancer genome atlas
TGF	transforming growth factor
TGF-β	transforming growth factor beta
TNM	tumor-node-metastasis
TP53	cellular tumor antigen p53
Tris	tris(hydroxymethyl)-aminomethane
TTK	dual specificity protein kinase TTK
WGS	whole-genome sequencing
WNT	wingless type
ZBTB2	zinc finger and BTB domain-containing protein 2

All gene names and symbols can be found in the HGNC database (http://www.genenames.org).

ABSTRACT

Colorectal cancer (CRC) is the third most common cancer, and the second most common cause of cancer mortality. Both somatic mutations and inherited genetic variation drive the development of CRC. Characterizing the underlying genetic changes is fundamental in basic cancer research. This knowledge may ultimately be translated into the development of more effective approaches for reducing cancer morbidity and mortality. The aim of this study was to gain novel insight into the molecular mechanisms behind CRC predisposition, as well as tumor progression and development.

Microsatellite instability (MSI) arises due to a defective mismatch repair system and is a feature of Lynch syndrome and a subset of all CRCs. MSI tumors are prone to repeat mutations, which in coding regions usually lead to premature termination codons (PTC). PTCs that occur in the end of the coding region of a gene might escape nonsense-mediated decay mechanisms. In the first project, we characterized all genes that were overexpressed in MSI CRCs and predicted to escape decay when mutated. The mitotic checkpoint kinase *TTK* was identified as a putative oncogenic target gene, with decay-escaping mutations in 59% (105/179) of the MSI CRCs screened. TTK is known to have an essential role in spindle assembly checkpoint (SAC) signaling; however, the mutated protein did not show SAC weakening. While no evidence of oncogenic mechanisms was observed, the high mutation frequency of *TTK* argues for biological significance.

Second, we sought to identify novel driver oncogenes with activating missense type changes in MSI CRCs. The exomes of 25 MSI tumors and respective healthy tissues were sequenced. A total of 15 candidate oncogenes with confirmed mutation hot spots were identified. Three genes, *ZBTB2, PSRC1* and *RANBP2*, displayed hot spot mutations also in the validation set of 86 MSI CRCs. Interestingly, the protein interactomes of ZBTB2 and PSRC1 consisted of many known cancer-related proteins and proteins with molecular functions relevant to cancer development and progression. In addition, the CRC-associated mutant form of ZBTB2 was shown to increase cell proliferation. Additional work is needed to further clarify the role of the identified somatic mutations in CRC tumorigenesis. Our results support the previous notion that CRC genomes are heterogeneous, characterized by a few frequently mutated genes, such as *BRAF* and *PIK3CA*, and a much larger number of genes mutated at intermediate frequencies, such as *HRAS* and here-identified *PSRC1, ZBTB2* and *RANBP2*. The candidate oncogenes identified in this thesis work might be used to develop personalized tumor profiling and therapy.

Inherited susceptibility is estimated to be involved in approximately one-third of all CRCs. However, few of these cases are associated with well-known highly penetrant mutations leading to inherited cancer syndromes. The great majority of inherited

CRC susceptibility remains still molecularly unexplained. A recent systematic sequencing study on CRC reported a set of somatically mutated genes, termed candidate cancer (CAN) genes. In study III, we examined the mutational profiles of 15 CAN genes for somatic mutations as well as for germline variants in 45 familial CRC cases. In our tumor set, six of the CAN genes were somatically mutated. In germline, three private missense variants were identified in *CSMD3, EPHB6* and *c10orf137*.

With novel sequencing tools at hand, another effort was performed with the aim to identify novel susceptibility genes for common familial CRC. In study IV, we sequenced the exomes of 96 independent cases with familial CRC. We focused our search on genes harboring rare putative loss-of-function (LoF) variants. In total, 11 novel candidate CRC susceptibility genes emerged from our efforts with putative LoF variants. These variants were absent or extremely rare in the general population. Seven loss-of-heterozygosity events, involving four genes, were observed in the data. In each occasion, the losses targeted the wild-type allele (P=0.0078), providing further support that true culprits are among the eleven genes. This study provides an interesting set of candidate predisposing genes, which might explain a subset of common familial CRC.

The germline variants identified in studies III and IV need to be validated in larger sample sets, representing different populations, to provide firm evidence for disease predisposition. Additional work is also needed to characterize the detailed functional and clinical relevance of the identified candidate CRC predisposing genes. This information, then, can ultimately be translated into tools for cancer prevention and early diagnosis of individuals carrying true predisposition alleles.

INTRODUCTION

Cancer refers to a large group of diseases, which may originate from most of the cell types and organs of the human body. The most common cancers are carcinomas, which are of epithelial origin (http://www.cancer.fi/syoparekisteri/en/, Finnish Cancer Registry, 2011 Statistics). All cancer cells share one important characteristic: they grow and proliferate in defiance of normal control. They may also acquire the capability to invade, disseminate from the site of the primary tumor and colonize distant organs. Tumors can be either benign (localized, noninvasive), which is the most common type, or malignant (invasive, metastatic). Metastases spawned by malignant tumors are the cause of nearly all cancer related deaths (Mehlen & Puisieux, 2006).

The development of cancer is a multistep process reflecting the accumulation of genetic and epigenetic alterations. These alterations drive the progression and transformation of cells from a normal to a more malignant state. The process in which tumors develop is analogous to that described in Darwinian natural selection. Alterations that increase the fitness of a neoplastic clone (cells with a common genotype) accumulate and result in clonal expansion. The fitness of a neoplastic cell is shaped by its interactions with other cells, soluble factors and the extracellular matrix in its immediate microenvironment (Merlo et al., 2006; Hanahan & Weinberg, 2000). Hanahan and Weinberg (2000) have described the following hallmark capabilities that a cell needs to acquire in order to reach a malignant state: sustained proliferative signaling, evasion of growth suppression, activation of invasion and metastasis, replicative immortality, induction of angiogenesis and resistance to cell death (Hanahan & Weinberg, 2000). Lately, two emerging hallmarks have been added to the list: reprogramming of energy metabolism and evading immune destruction. In addition, genomic instability and inflammation have been proposed as "enabling characteristics" that facilitate the acquisition of the above mentioned hallmarks (Hanahan & Weinberg, 2011).

Cancer is generally a slowly progressing disease and the development of a clinically detectable solid tumor is estimated to take up to several decades (Loeb et al., 2003). The risk of developing cancer is influenced by environmental and lifestyle factors, as well as by the set of genomic variants present in the germline of an individual. Some of the most common lifestyle and environmental risk factors for cancer are smoking, diet and obesity. Moreover, infectious agents are estimated to cause approximately 15 % of all cancers. Well-known examples are *Helicobacter pylori* in gastric cancer (Parsonnet et al., 1991) and human papillomaviruses in cervical cancer (Hausen & de Villiers, 1994).

REVIEW OF THE LITERATURE

1 Cancer as a genetic disease

It is now widely accepted that all cancers arise as a result of numerous alterations that have occurred in the DNA sequence of cancer cells. These sequence variants can be transmitted through the germline of an individual and result is cancer susceptibility or they can be somatically acquired mutations. The germline variants are present in the fertilized egg from which the individual develops and will thus be present in all the cells of the human body. Somatically acquired mutations, including base substitutions, insertions and deletions of bases, rearrangements and copy number alterations, occur in the genomes of cells upon mitotic cell division (Stratton et al., 2009). Additional mutations accumulate when cells divide further and only when several genes are defective, cancer will develop. It has been suggested that the great majority of cancers arise when two to eight sequential alterations have occurred, during several decades, in genes with functions relevant to cancer (Vogelstein et al., 2013) (Figure 1).

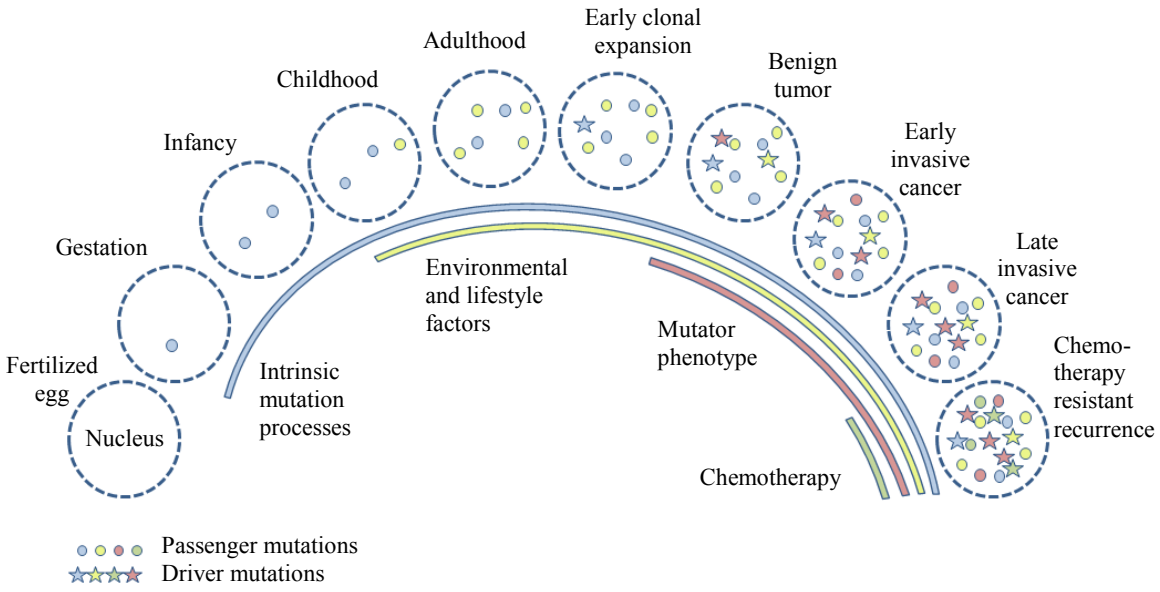

Figure 1. Somatic mutations accumulate in a cell that will form a neoplastic tumor cell colony of a malignant cancer. A malignant cancer cell develops via a lineage of mitotic cell division from the fertilized egg. Somatic mutations (represented by colored symbols) accumulate over a lifetime and this process is affected by both intrinsic and environmental factors. A subset of these mutations are driver mutations, which confer selective growth advantage upon the neoplastic clone, while the great majority are neutral passenger mutations. The figure was drawn based on Stratton et al., 2009.

Somatic mutations occur in every cell division, at a rate of approximately 10×10^{-7}, in a more or less random fashion (Araten et al., 2005). In a neoplastic clone that is to

become a cancer, a subset of mutations has by chance occurred in genes essential for tumor development. Mutations in these genes, also called cancer genes confer selective growth advantage for the neoplastic clone, which then undergoes clonal expansion. Such driver mutations enable the cells to acquire hallmark capabilities, such as resistance to cell death or evasion of growth suppression. These capabilities are required for metastatic cancer to develop. There are also numerous passenger mutations in the final clonal expansion that do not confer selective growth advantage and are biologically neutral. These mutations were by chance present in the progenitor cell that later underwent clonal expansion (Greenman et al., 2007; Hanahan & Weinberg, 2000) (Figure 1).

Other important factors that regulate tumorigenesis, in addition to DNA sequence alterations, are epigenetic alterations and microRNAs (miRNAs). The epigenome undergoes several alterations during tumor progression, such as genome-wide loss of DNA methylation (hypomethylation) and excessive promoter methylation at CpG islands (hypermethylation) (Shen & Laird, 2013). miRNAs are small non-coding RNAs of 20-22 nucleotides, which are typically differentially expressed in cancers and can alter the expression of cancer genes (Croce & Calin, 2005).

1.1 General features of cancer genes

Of all the cancer genes known to date, approximately 90% show somatic mutations, 20% show germline mutations and 10% show both. The most common mutation types in these genes are chromosomal translocations, frequently seen in lymphomas, leukemias and sarcomas (Futreal et al., 2004). Cancer genes have classically been divided into oncogenes and tumor suppressor genes depending on their mutation patterns and the effect of the mutations on gene function and cellular processes. These classifications may be arbitrary and oversimplified, however, they facilitate certain molecular genetic analyses and the detection of specific mutation patterns (Vogelstein & Kinzler, 2004).

1.1.1 Oncogenes

Oncogenes are altered in cancers in ways that render the gene constitutively active or active under situations when the wild-type is not. On cellular level these alterations act in a dominant manner, meaning that one allele is usually sufficient to confer a selective growth advantage to the cell. The normal equivalents of oncogenes are called proto-oncogenes, and proteins encoded by these genes usually function as transcription factors, growth factors, signal transducers or apoptotic regulators. These proteins positively regulate cellular processes such as cell growth, survival and migration. When a proto-oncogene becomes activated by intragenic mutation, chromosomal translocation or gene amplification, it transforms into an oncogene and might contribute to the initiation and progression of cancer (Croce, 2008). Since

the identification of the first human oncogene *HRAS*, with a glycine to valine substitutions at codon 12 in the human bladder carcinoma cell line T24/EJ (Reddy et al., 1982), several human oncogenes have been discovered (Croce, 2008).

Oncogenes are frequently activated by intragenic mutations. The patterns of mutations tend to be highly nonrandom, with most of the mutations enriched in certain regions of the protein. It has been estimated that typical oncogenes have > 20% of missense mutations in recurrent positions (Vogelstein et al., 2013). The most commonly mutated oncogenes in human cancers are the *RAS* genes (*KRAS, HRAS* and *NRAS*), which code for small GTPases that are involved in transmitting signals within the cell. Oncogenic RAS mutations result in constitutive mitogenic signaling, one of the most fundamental trait of cancer cells (Pylayeva-Gupta et al., 2011). BRAF, acting downstream of RAS in the MAPK/ERK pathway, also shows activating mutations in many cancers, most commonly at codon V600. This residue is within the activating loop of the kinase domain and constitutively activates the enzyme. The activated kinase phosphorylates downstream targets, such as extracellular signal-regulated kinase (ERK), which ultimately leads to aberrant cell growth (Wan et al., 2004). Oncogenes can also be activated by chromosomal translocations, such as *MYC* in Burkitt's lymphoma and *BCR-ABL* in Chronic Myelogenous Leukemia, or through gene amplification as often seen for *MYC, EGFR* and *ERBB2* in several different cancers (Croce, 2008).

1.1.2 Tumor-suppressor genes

In normal cells, tumor-suppressor genes often function to restrain cell growth and division and to stimulate cell death. In cancer, these genes are frequently altered leading to loss of function or reduction in protein activity. Tumor-suppressor genes are recessive in nature: mutations in both alleles are generally required to confer a selective growth advantage to the cell. This principle is known as the "two-hit" hypothesis and was first proposed by Alfred Knudson (1971). According to this model, familial form of cancers may arise by two inactivating alterations of which one is inherited through the germline and the other is acquired somatically. Conversely, sporadic cancers require two somatically acquired hits and thus such cancers usually develop at a later age (Knudson, 1971). The inherited inactivated allele tends to show small intragenic mutations, whereas the remaining allele is usually inactivated by similar mutations or by loss of heterozygosity (LOH), caused by for instance mitotic recombination (Knudson, 2002).

The *RB1* gene is an example of a classical tumor suppressor gene (also known as a gatekeeper) that drives cell progression in a direct manner when both alleles are inactivated and predisposes to tumors of the retina (Friend et al., 1986; Kinzler & Vogelstein, 1997). RB1 is a critical regulator of cell-cycle progression and when inactivated leads to persistent cell proliferation and evasion of growth suppression

(Hanahan & Weinberg, 2011). The tumor suppressor TP53 is another key control node that regulates cell-cycle progression. The *TP53* gene is mutated in half of all human cancers and the rest of the cancers often have alterations in its interaction partners. Unlike RB1, TP53 receives signals from within the cell upon several forms of cellular stress, such as hypoxia and DNA damage. Inactivated TP53 leads to resistance to programmed cell death (apoptosis) and evasion of growth suppression. Patients with Li-Fraumeni syndrome have germline mutations in *TP53* (Vogelstein et al., 2000; Prives, 1998; Hanahan & Weinberg, 2011). Other well-known classical tumor suppressors are *APC* (Levy et al., 1994) in CRC, and *BRCA1* (Miki et al., 1994) and *BRCA2* in breast cancer (Wooster et al., 1995).

There are exceptions to the classical two-hit hypothesis when a mutation or loss of a single-copy of a tumor suppressor gene plays a significant role in tumorigenesis. In some occasions, a single-copy event may be preferentially selected for in tumor evolution, instead of biallelic inactivation that might lead to cell death or senescence. The term haploinsufficiency refers to the scenario when inactivation of a single allele is enough for aberrant protein function and promotion of cancer (Santarosa & Ashworth, 2004). One example is the haploinsufficient loss of PTEN that can provide growth advantage, while avoiding senescent signals of TP53 that a complete loss of PTEN would induce. Another exception to the classical two-hit hypothesis is when a single-copy mutation functions in a dominant negative manner, interfering with the normal protein produced by the remaining wild-type allele (Berger et al., 2011).

A subclass of tumor suppressor genes are the stability genes (also called caretakers). These genes promote tumorigenesis indirectly by creating genomic instability. Normally these genes function to keep the number of genetic alterations low but upon their inactivation the mutation rate in all other genes is increased. However, only mutations that target oncogenes or tumor suppressor genes will be preferentially selected for and have a tumor promoting effect. Similar to classical tumor suppressor genes, both alleles are generally inactivated in the tumor. Stability genes include the mismatch repair, nucleotide-excision repair and base-excision repair genes. Also genes involved in mitotic recombination and chromosomal segregation belong to this class, for example *BRCA1* and *ATM* (Vogelstein & Kinzler, 2004; Kinzler & Vogelstein, 1997).

1.2 Inherited predisposition to cancer

The great majority of common cancers arise sporadically and are highly influenced by environmental and lifestyle factors. An estimated 5-10% of all cancers are inherited, due to highly penetrant germline mutations that cause rare inherited cancer syndromes. Another 15-20% of all common cancers are known as “familial”, which can be defined as clustering of cancer in a family more frequently than

expected (Nagy et al., 2004). Still today, the molecular background of "familial" cancers remains largely unexplained. The familial clustering is most likely due to the inheritance of common low-penetrance alleles and rare moderate-penetrance alleles, as well as epistatic interactions (Fletcher & Houlston, 2010). Research has lately focused on identifying novel predisposing variants behind familial forms of cancer. However, challenges arise due to the multifactorial nature of the disease, related to the heterogeneity observed on both cellular and genetic level. Identification of novel susceptibility genes is important, not only to gain better understanding of cancer biology in general but also for the identification of novel targets for therapeutic interventions. Also, identifying individuals at increased risk is of immediate clinical relevance.

1.2.1 Inherited cancer syndromes

A small fraction of common cancers can be explained by high-penetrance germline mutations that cause hereditary cancer syndromes with often quite distinct clinical features. There are several characteristics of hereditary cancers, such as multiple affected individuals in the family over several generations, early age of onset, and multiple primary cancers in one individual. Many of the known cancer syndromes show complete penetrance by the age of 70. However, due to factors such as phenotypic variability and age-related penetrance, some families with an inherited cancer syndrome do not show the above mentioned characteristics (Nagy et al., 2004). Predisposing alleles underlying rare hereditary cancer syndromes usually have a minor allele frequency (MAF) less than 0.1% and confer high-risk with odds ratio >10. However, on population level they confer a small attributable risk (Fletcher & Houlston, 2010).

To date, more than 100 genes have been reported to cause Mendelian inherited cancer syndromes. Most syndromes fit autosomal dominant model with defects in tumor suppressor genes that conform to the two-hit model of cancer susceptibility. However, there are also syndromes that are of autosomal recessive nature, usually resulting from defects in stability genes (Cazier & Tomlinson, 2010). Classical genetic linkage-analysis and positional cloning has led to the discovery of many highly penetrant genes for common cancers. This was successfully performed for genes such as *BRCA1* and *BRCA2* (Hall et al., 1990; Wooster et al., 1995) in breast and ovarian cancer, *APC* (Bodmer et al., 1987; Nishisho et al., 1991) and mismatch repair genes (Peltomäki et al., 1993; Lindblom et al., 1993) in CRC, and *CDNK2A* (Cannon-Albright et al., 1992; Piepkorn, 2000) in melanoma.

1.2.2 Other forms of cancer-predisposing variation

Common cancers are known to cluster in families, and individuals with a first-degree relative affected have a two-to-four-fold higher risk of developing cancer

(Goldgar et al., 1994; Johns & Houlston, 2001). Also, most common cancers show higher concordance in monozygotic twins than in dizygotic twins. Heritability has been estimated to account for 42, 35 and 27 % of the variation in susceptibility to prostate, colorectal, and breast cancer, respectively (Lichtenstein et al., 2000). Most known cancer predisposing genes cause Mendelian inherited cancer syndromes, and explain only a small part of the entire heritable fraction of common cancers. This has led researchers to question where the "missing heritability" can be found. Potential sources of "missing heritability" could be variants of low MAF (0.5% < MAF < 5%) or of rare variants (MAF < 0.5%). Another source might be structural variation, including copy number variants and copy neutral variation, such as translocations. (Manolio et al., 2009).

The "rare variant hypothesis" proposes that a large fraction of the inherited susceptibility may be due to the summation of rare moderately penetrant risk alleles (with MAF ≤2 % and odds ratio ≥ 2) that each act independently and dominantly. These are thought to be mostly population specific due to founder effects that have resulted from genetic drift. Both next generation sequencing (NGS) and candidate gene sequencing approaches are thought to enable the identification of such variants (Bodmer & Bonilla, 2008; Bodmer & Tomlinson, 2010). To date, only few robustly validated moderate-penetrance genes have been identified in common cancers, such as *CHEK2* (Meijers-Heijboer et al., 2002; Vahteristo et al., 2002) in breast cancer and *MUTYH* (Al-Tassan et al., 2002) in CRC.

The "common disease-common variant" model proposes that alleles of high frequency (MAF >10 %) and low penetrance (typically odds ratio <1.5) contribute to the susceptibility of common cancers. Several common risk loci have successfully been identified for many common cancers by genome-wide association (GWA) studies. However, pinpointing the disease-causing variant at these risk loci has proven difficult (Fletcher & Houlston, 2010). The "rare variant hypothesis" and "common disease-common variant" models are contradictory, and a more continuous and comprehensive approach is more likely to model the true underlying genetic predisposition.

It is also argued that much of the remaining inherited susceptibility can be explained by the co-inheritance of several genetic variants, known as the "polygenic model of predisposition". Each individual is thought to carry a handful of variants of low/moderate-risk that exist in varying frequency in the population. An individual at very low risk might carry mainly low-risk alleles, whereas a person at higher risk might have one or more moderate-risk alleles (Fletcher & Houlston, 2010). It has been hypothesized that the polygenic basis of common diseases might be manifested in the regulation or function of one or more signaling pathways. Genetic variation at several different loci could cause many slight changes that together result in deregulation of key cellular signaling pathways (Sullivan et al., 2012).

2 The era of large-scale genome sequencing

In 2008, the first human cancer genome was sequenced by using NGS, where thousands to millions of DNA templates are processed in parallel (Ley et al., 2008; Shendure & Ji, 2008). Today, the cost of NGS has reduced more than 100-fold since the first genomes were sequenced. In basic and clinical research, it is now routine to sequence several exomes (i.e., the coding regions of the genome) and whole genomes accurately and rapidly. Over the next few years several thousand more genomes will be sequenced. Also, it is estimated that, as the costs drop even further, routine NGS sequencing will become part of every clinic. This vast amount of data will provide us with a detailed picture of the underlying inherited variations and acquired somatic mutations that drive tumor development and progression. However, challenges emerge related to interpretation of NGS data in meaningful terms. Further progress in this area will require carefully designed studies that are optimized to detect causal variants. Ultimately, this data will provide considerable increase in the knowledge of cancer biology and potentially novel opportunities for the development of new cancer treatments (Vogelstein et al., 2013; Kilpivaara & Aaltonen, 2013).

2.1 Human genomic variation

As a prerequisite for understanding how different germline variants contribute to cancer risk, we need to understand the spectrum of allelic variation in healthy individuals. This is particularly the case for population-specific rare variants that are thought to be enriched for disease susceptible variants (MacArthur et al., 2012). To date, several large-scale sequencing studies on human genomic variation have been performed, for example studies that are part of the 1000 Genomes Project. The 1000 Genomes Project is an effort where 1,092 individuals from 14 populations (including 93 individuals from Finland) have been low-coverage whole-genome and exome sequenced (1000 Genomes Project Consortium et al., 2012). The data provide researchers with a comprehensive resource on human genomic variation.

It has been reported that every individual carries approximately 2,500 non-synonymous variants at conserved sites and as many as 150 LoF variants (stop-gains, framesifting indels or splice-site variants). Most of the LoF variants are common (MAF >5%) or low-frequency (MAF 0.5-5%) with the number of rare LoF variants (MAF <0.5%) being much lower, approximately 10-20 per individual (1000 Genomes Project Consortium et al., 2012). Human genomic variation shows substantial population differences, especially for variations that are rare. More than half of all the rare variants found in the 1000 Genomes Project were found in a single population (Gravel et al., 2011; 1000 Genomes Project Consortium et al., 2012). These results highlight the challenge to replicate disease-association for rare variants in different populations and the challenge to find causal variants from the large

number of neutral background variation.

2.2 Novel insights into cancer predisposition

Until now, approaches for detecting rare/low frequency coding variants of moderate penetrance for common cancers have been poor. Attractive patient groups to search for such variants are common familial cases, with few affected first-degree relatives and early-diagnosed cancer patients. Common cancer families are usually too small for linkage analysis, and the variants are too rare to be detected in GWA studies. In addition, candidate gene screens have been heavily biased towards genes with previous supporting functional or genetic data (Bamshad et al., 2011). NGS, including exome and whole genome sequencing (WGS), is a powerful new tool to examine the underlying genetic architecture of common cancers in an unbiased and systematic manner (Figure 2).

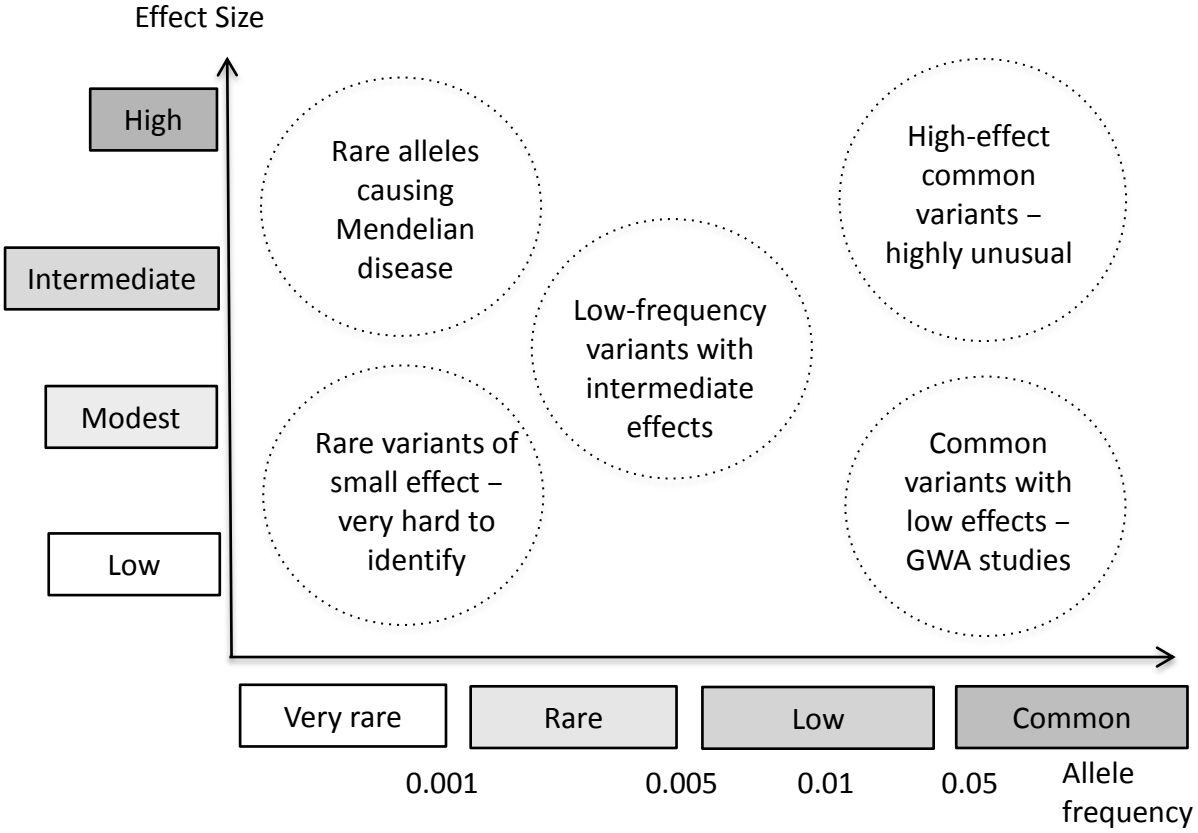

Figure 2. Genetic architecture of cancer risk. Genetic variants in the population can be placed on a continuum of allele frequency and effect size. Mendelian syndromes occupy the upper left circle, consisting of rare high-penetrance alleles mostly identified by linkage-analysis. GWA studies have proven successful in identifying common variants of low effect size (lower right). The middle, which consists of rare/low-frequnecy variants of varying effects have been fairly unexplored. Advances in sequencing technologies allow for the exploration of the relationship between such variants and cancer predisposition (figure drawn based on McCarthy et al., 2008).

Although NGS is considered a highly attractive approach, there are still challenges related to data interpretation. A key challenge is how to pinpoint key susceptibility alleles among a large number of non-pathogenic background variations and sequence artifacts. Also, optimal NGS study designs need to take into account variables, which include: inheritance pattern, population structure and the extent of locus heterogeneity. Such variables affect, for example, the sample size required to obtain sufficient power to detect robust disease-association. Often statistically weak associations need further support from additional information related to, for example, preferential selection of the locus in the tumor tissue or additional functional evidence (Bamshad et al., 2011; Bansal et al., 2010).

To date, there are fairly few examples where NGS has been successfully utilized to identify novel cancer predisposing loci for common cancers. Several studies including small sample sets have reported novel predisposing cancer genes that have subsequently failed robust validation in other sample materials, for example *PALB2* (Jones et al., 2009) and *ATM* (Roberts et al., 2012) in pancreatic cancer (Grant et al., 2013). WGS was successfully utilized in a study conducted in Iceland, where they identified a novel rare single-nucleotide variant at 8q24 that predisposes individuals to prostate cancer. The association of the rare variant was confirmed in other European populations, and it was shown to confer a slight increase in prostate cancer risk; however, the risk was higher (odds ratio = 2.90) compared to those variants identified previously by GWA studies (typically with odds ratios < 1.5) (Gudmundsson et al., 2012). In the near future, these studies will most likely be performed in a similar fashion to GWA studies, with very large sample sizes that allow for sufficient statistical evidence to pinpoint true predisposing variants based on the association evidence alone. In the meantime, it is important to optimize study design and data analysis strategies to detect pathogenic variants in smaller sample sets.

2.3 Cancer genome landscapes

Over the last years, comprehensive large-scale sequencing efforts have revealed new insights into the cancer genome landscapes of many common cancers. One of the largest ongoing efforts is that conducted by the Cancer Genome Atlas project, were 20 “mutatomes” from different cancers are being profiled (Cancer Genome Atlas Network, 2012).

The average number of somatically acquired alterations in a particular tumor largely depends on the tumor type, with most common solid tumors showing an average of 33 to 66 non-synonymous somatic mutations. Outliers are melanoma and lung cancers, with a high number of mutations, and pediatric tumors and leukemias with a low number of mutations (Vogelstein et al., 2013) In addition, tumors with a DNA repair defect represent another group of outliers that have up to 1000 non-

synonymous mutations per tumor (Palles et al., 2013; Cancer Genome Atlas Network, 2012). Recent efforts have also highlighted the fact that most somatic mutations in a given tumor type are passenger mutations and do not confer any selective growth advantage upon the cell. How to find the true driver genes in the full repertoire of somatic mutations is still a challenge; however, several prioritizing strategies have been proposed related to mutation frequency, gene length, gene mutation patterns and other parameters (Vogelstein et al., 2013).

For most cancer types, there are a few genes that are mutated at high frequency and a much larger number of genes mutated infrequently. The genomic landscape of common cancers has thus revealed a similar topography of mountains and hills. Studies have shown that two tumors of the same histopathologic subtype are fairly distinct in respect to their genetic alterations (Sjöblom et al., 2006; Wood et al., 2007). Vogelstein *et al.* (2013) recently highlighted the need for better understanding of altered signaling pathways rather than individual genes. They proposed that all of the cancer genes can be classified into one or more of 12 pathways, and these pathways can be further organized into three core cellular processes: cell fate (for example *APC* and *NOTCH*), cell survival (for example *RAS* and *PIK3CA*) and genome maintenance (for example *TP53* and *MLH1*) (Vogelstein et al., 2013).

3 Colorectal cancer

3.1 Introduction to colorectal cancer

CRC is still one of the leading types of cancer. Worldwide, it's the fourth most common cancer in men and the third in women. There is significant international variation in incidence rates; North America and Europe have high rates, and Asia, Africa and South America have low rates (Center et al., 2009). In Finland, the incidence is 27.9 per 100,000 in males and 19.4 per 100,000 in females; with approximately 2800 new cases diagnosed each year. According to the Finnish cancer registry data, the 5-year survival rate is around 60-65% for all cases (http://www.cancer.fi/syoparekisteri/en/). The lifetime risk of CRC in the general population is approximately 5-6 % (Jemal et al., 2008).

CRC is a complex disease influenced by both genetic and environmental factors. The genetic risk factors will be described in detail in the next chapters. Lifestyle and environmental risk factors include, for instance, diet, physical inactivity and smoking (Giovannucci, 2002; Botteri et al., 2008). Interestingly, physical inactivity has been estimated to cause up to 10% of the burden of CRC (Lee et al., 2012). Environmental and lifestyle factors partly explain the high rate of CRC observed in the Western world. In addition, an increased risk for CRC has also been reported for individuals with inflammatory bowel disease (Dyson & Rutter, 2012). There are

factors that reduce CRC risk; one well established example is aspirin, which has been shown to reduce CRC risk and improve survival after diagnosis (Chia et al., 2012).

There are two widely used staging systems when diagnosing CRC; the TNM (tumor, node, metastasize) staging system and the Dukes Classification (Compton & Greene, 2004) (Table 1; modified from Union for International Cancer Control, http://www.uicc.org). Tumors of TNM stage I or II, which are local invasive cancers, can often be cured by surgical removal. Stage III tumors, which have spread to regional lymph nodes, are curable by surgery combined with adjuvant therapy in around 73 % of cases. Cancers that have metastasized (stage IV) are often fatal; however, improvements in anti-angiogenic therapy and EGFR based therapy have improved patient survival (Heinemann et al., 2013). Early detection of CRC has a crucial impact on survival. For patients with stage A disease, according to Dukes classification, the 5-year overall survival rate is as high as 95 %, but only 0-7 % for Dukes D stage patients (Weitz et al., 2005). Accurate cancer staging is important not only for appropriate evaluation of therapies, prediction of survival and prognosis, but also for cancer research in general.

Table 1. TNM staging system and Dukes classification of CRC

Stage	Definition
T0	No evidence of primary tumor
Tis	Carcinoma in situ: intrepithelial or intramucosal
T1	Tumor invasion into submucosa
T2	Tumor invasion into muscularis propria
T3	Tumor invasion through muscularis propria
T4	Tumor invasion into other organs or through visceral peritoneum
N0	No evidence of regional lymp node metastasis
N1	Metastasis into 1-3 regional lymph nodes
N2	Metastasis into ≥4 regional lymph nodes
M0	No evidence of distant metastasis
M1	Distant metastasis

Stage Grouping	Dukes stage	5-year survival (%)*
Stage I: T1-2, N0, M0	Dukes A	80-95
Stage II: T3-4, N0, M0	Dukes B	65-75
Stage III: Any T, N1-2, M0	Dukes C	25-60
Stage IV: Any T, Any N, M1	Dukes D	0-7

* Weitz et al., 2005

3.2 Colorectal tumorigenesis

CRC develops from rapidly renewing epithelial cells lining the colon or rectum of the gastrointestinal tract. The epithelial cells form a single sheet with crypts protruding into the underlying connective tissue (Humphries & Wright, 2008). Stem cells are located at the base of the crypt, forming the stem-cell niche together with mesenchymal cells. The stem cells have the capability to regenerate all colonic cell types. In normal conditions, the epithelial stem cells receive homestatic signals from the surrounding mesenchymal myofibroblasts, including WNT-signaling ligands (Fevr et al., 2007). It is thought that the initial mutational event in CRC occurs in the epithelial stem cells, which then come to dominate the stem-cell niche through clonal expansion. The cells migrate up the crypt, fail to differentiate normally, and finally spread into the colonic epithelium (Humphries & Wright, 2008).

It is now widely appreciated that CRC results from the accumulation of genetic and epigenetic alterations, which lead to the transformation of normal colonic epithelium to colorectal adenocarcinoma. The development of colorectal adenocarcinoma is characterized by a series of well-defined histopathological changes, each of which is accompanied by specific genetic alterations (Hanahan & Weinberg, 2000; Fearon & Vogelstein, 1990). A key feature underlying CRC development is genomic instability, which leads to the acquisition of multiple genetic alterations that then drive malignant transformation (Loeb, 1991; Fodde et al., 2001). It is thought that genomic instability occurs early in the tumorigenesis process, already during the initiation of adenoma formation (Shih et al., 2001; Nowak et al., 2002).

CRC cells can acquire increased mutability of their genomes through several different molecular pathways. CRC tumors are usually divided into those with chromosomal instability (CIN) and those with microsatellite instability (MSI) (Aaltonen et al., 1993; Kinzler & Vogelstein, 1996) (Figure 3). Both of these pathways are effective mechanisms to remodel the genome in ways that favor evolution towards neoplasia. More recently, tumors have been subcategorized based on their mutation rate. The TCGA study recently described CRCs to be either non-hypermutated or hypermutated based on the number of mutations on nucleotide level. Non-hypermutated cancers, which represent the large majority of CRCs (84%), are usually microsatellite stable (MSS) and show CIN (Cancer Genome Atlas Network, 2012).

3.2.1 Chromosomal instability

The CIN pathway reflects the classical adenoma-carcinoma sequence, the progressive accumulation of point mutations in genes such as *APC, KRAS* and *TP53*, in addition to frequent chromosomal losses and gains, especially losses on

chromosome arms 5q, 17p and 18q (Vogelstein et al., 1988; Fearon & Vogelstein, 1990) (Figure 3). CIN is thought to arise at the very first steps of colorectal tumorigenesis, already in abberant crypt foci (ACF). ACFs develop before colorectal polyps and are the earliest detectable change of the adenoma-carcinoma sequence (Luo et al., 2006; Vogelstein et al., 1988).

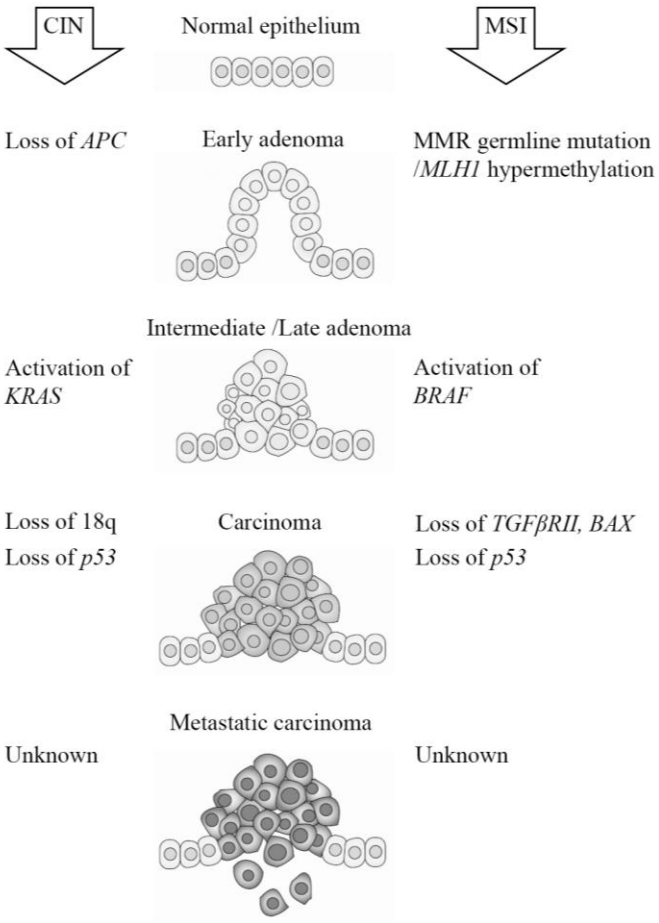

Figure 3. The stepwise progression of CRC. The main genetic alterations that drive tumorigenesis in both CIN and MSI tumors are shown. The schematic figure is modified from Knudson, 2001. See text for references.

Biallelic inactivation of *APC* at 5q is the earliest mutational event observed in the adenoma-carcinoma sequence and seems to be required for the initiation of clonal evolution (Powell et al., 1992). Approximately 70-80% of sporadic CRCs show somatic inactivation of *APC* (Kinzler & Vogelstein, 1996; Polakis, 2007). A small subclass of tumors with wild-type *APC* shows mutations in other members of the WNT pathway, such as *CTNNB1* (also known as *β-catenin*) (Morin et al., 1997). *APC* mutations can be found already at ACFs and are tightly associated with the degree of dysplasia of these lesions (Jen et al., 1994; Smith et al., 1992). The crypts in which the *APC*-mutant cells lie become slowly dysplastic as abnormal cells start to accumulate. Whether *APC* mutations occur on a background of genetic instability or triggers genetic instability remains an open question. Evidence suggests that *APC* is

mutated when cells are near-diploid rather than aneuploid (Michor et al., 2005; Fodde et al., 2001). Inactivation of APC seems to underlie both tumor initiation and promotion, since APC has also been reported to directly enhance mutation rates through chromosomal instability (Fodde et al., 2001).

Additional mutations, such as activating mutations in *KRAS*, are required for adenoma growth and progression. Approximately 40% of CRCs show *KRAS* mutations with most mutations affecting codons 12 and 13 (Fearon & Vogelstein, 1990; Wood et al., 2007; Vogelstein et al., 1988). Oncogenic KRAS has been shown to contribute to tumor progression at an early stage, during transition from intermediate to late adenoma (Lamlum et al., 2000). In *KRAS* wild-type tumors, the RAF–MAPK pathway might be activated by mutations in *NRAS, EGFR* (*ERBB1*) or *ERBB2* (*HER2*) (Cancer Genome Atlas Network, 2012).

For the polyps to progress into cancer, additional mutational events are required, such as loss of chromosome 17p, which is found in more than 75% of all CRCs (Rodrigues et al., 1990). The *TP53* gene is thought to be the main target of 17p loss with somatic mutations, mostly missense mutations, frequently affecting the remaining *TP53* allele. The inactivation of *TP53* often coincides with transition of large adenomas into invasive carcinomas (Baker et al., 1990). Loss of 18q is another frequent event observed in CRCs. The genes that underlie molecular pathology are thought to be *SMAD2* and *SMAD4*, mutated in a fraction of CRCs (Wood et al., 2007; Leary et al., 2008).

It is estimated that the entire process from ACFs to invasive carcinomas takes between 20-40 years. During this period, there is a constant increase in CIN (Rajagopalan et al., 2003). The molecular basis behind CIN remains largely unexplained. It is thought that genes that regulate the formation of the mitotic spindle and proper alignment and segregation of chromosomes at mitosis may contribute to CIN (Grady, 2004; Barber et al., 2008), such as *BUB1, MAD2* and *APC* (Cahill et al., 1998; Alberici & Fodde, 2006).

3.2.2 Microsatellite instability

A subset of CRC cancers have hypermutated genomes and show a so called "mutator phenotype", due to defects in genes that function in the maintenance of genomic stability. These cancers are fairly stable on chromosomal level, with near-diploid genomes; however, they show high mutation rates on nucleotide level. Hypermutated tumors have mutation rates of 10-100 per 10^6 bases, whereas non-hypermutated tumors show mutation rates of less than 10 per 10^6 bases. The great majority of hypermutated tumors show microsatellite instability, driven by a defective mismatch repair system (Cancer Genome Atlas Network, 2012; Loeb, 1991). There are differences in the sequence of genetic events observed in hypermutated

versus non-hypermutated CRCs, which might imply that they undergo distinct pathways to tumorigenesis. Hypermutated CRCs generally show fewer mutations in *APC*, *KRAS* and *TP53* and higher mutation frequencies in *BRAF* and TGF-beta pathway related genes (Cancer Genome Atlas Network, 2012; Jass, 2004) (Figure 3).

Approximately 15% of CRCs develop through the microsatellite instability (MSI) pathway, which is driven by defects in the mismatch repair system. The defect can be inherited, which is the case in Lynch syndrome, or acquired, as in sporadic MSI tumors. In patients with Lynch syndrome, the MSI phenotype is caused by germline mutations in mismatch repair genes (mostly *MLH1* and *MSH2*) (Aaltonen et al., 1993; Ionov et al., 1993; Thibodeau et al., 1993). Sporadic MSI CRCs are typically caused by epigenetic silencing of the *MLH1* gene (Kane et al., 1997; Veigl et al., 1998). Previous studies have shown patients with MSI tumors to have better prognosis and a lower risk of recurrence than other CRCs (Watanabe et al., 2001; Van Schaeybroeck et al., 2011). MSI tumors differ genetically and clinicopathologically from the rest of the CRC tumors. Common features of MSI tumors are proximal location, lymphocytic infiltration, poor differentiation and mucinous features (Vilar & Gruber, 2010).

3.2.2.1 The mismatch repair system

Microsatellites are repeated-sequence motifs, consisting of simple mono-, di-, tri- and tetranucleotide DNA repeats, found all across the genome in large numbers (Ellegren, 2004). These sequences are prone to mutations. Due to replication strand slippage, the DNA polymerase occasionally stutters while copying microsatellites, leading to longer or shorter versions of the repeats in the newly synthetized strand. These replication errors can be recognized and corrected by the MMR system. Base mismatches made by the DNA polymerase may also be erased by MMR proteins. The predominant components of the MMR machinery are MutSα, MutSβ and MutLα. First, MutSα (a heterodimer of MSH2 and MSH6) or MutSβ (a heterodimer of MSH2 and MSH3) locates the mismatch or the insertion-deletion loop. Second, MutLα (a heterodimer of MLH1 and PMS2) forms a complex with MutSα or MutSβ to subsequently activate the repair process (Boyer & Farber, 1998; Jiricny, 2006).

3.2.2.2 Microsatellite instability target genes

In cells with MMR defects, mismatches remain uncorrected, which ultimately results in a mutator phenotype. The great majority of the mutations in MSI tumors are passenger events with no effect on malignant growth. Occasionally, frameshift mutations that result in protein truncation, or other alterations in the protein product, target a crucial gene and provides the cell with a growth advantage (Loeb, 1991; Boland et al., 1998). To distinguish real driver MSI target genes from passengers is challenging. Studies on non-coding repeats have revealed the background mutation frequency to be surprisingly high in MSI CRCs, with a strong

correlation to repeat type and length (Sammalkorpi et al., 2007). Several criteria have been suggested for the identification of real MSI target genes, such as high mutation frequency, biallelic inactivation, mutation in MSS cancers and supporting functional evidence (Boland et al., 1998). Examples of well-established target genes, with high mutation frequencies and robust functional evidence, are *TGFBR2* (Markowitz et al., 1995; Wang et al., 1995) and *BAX* (Rampino et al., 1997; Ionov et al., 2000).

Frameshift mutations generally result in premature termination codons (PTC) and a truncated protein. For this reason the great majority of MSI target genes are thought to show loss of function effects. Translation of aberrant transcripts is usually inhibited by the nonsense-mediated decay (NMD) system that degrades mRNAs containing PTCs (Isken & Maquat, 2007). However, aberrant transcripts may escape the NMD-system, typically those with PTCs located at the very end of the mRNAs (Nagy & Maquat, 1998).

3.2.3 The ultramutated phenotype

Two recent studies identified a small novel class of hypermutated CRCs that result from exonuclease domain mutation (EDM) in *POLE* and *POLD1* (Palles et al., 2013; Cancer Genome Atlas Network, 2012). POLE and POLD1 form the catalytic and proofreading subunits of the two central polymerases ε and δ, which replicate DNA (Nick McElhinny et al., 2008). The mutations can be inherited and lead to a rare condition termed polymerase proofreading-associated polyposis (PPAP). Affected individuals with such a condition have a high risk of multiple colorectal adenomas and carcinomas. Somatic mutations in *POLE* have been reported in CRCs as well as endometrial cancer. Currently, there is no proper evidence for the existence of pathogenic somatic POLD1 mutations. Both germline and somatic EDM mutations result in an "ultramutated" phenotype, with mutation rates of over 50 per 10^6 bases. Current evidence suggest these tumors to be of MSS type (Palles et al., 2013; Cancer Genome Atlas Network, 2012).

3.2.4 Altered signaling pathways in colorectal cancer

WNT signaling is a central pathway in embryogenesis and colonic homeostasis in the adult (Lin et al., 2008). In colorectal tumorigenesis, the initiating event is thought to be the activation of the WNT signaling pathway (Powell et al., 1992). In normal cells and in the absence of WNT ligand, APC associates with axin, glycogen synthase kinase 3β (GSK-3β) and casein kinase 1 (CK1) to form a so-called β-catenin destruction complex. β-catenin is phosphorylated by this complex, resulting in β-catenin ubiquitylation and subsequent proteosomal degradation (Polakis, 2002). However, in cells with mutations in members of the WNT signaling pathway, β-catenin accumulates and translocates to the nucleus. Once in the nucleus, it interacts with the T-cell factor/lymphoid enhancer factor (TCF/LEF) family of transcription

factors to activate specific WNT target genes (Tetsu & McCormick, 1999), such as *MYC* and *CCND1* (previous name *cyclin D1*) (Polakis, 2007; He et al., 1998).

Another early event in CRC is increased signaling through the RAS/RAF/MEK/ERK pathway. The pathway is known to regulate proliferation and differentiation and is activated by stimuli such as growth factors and cytokines. Both activating *KRAS* and *BRAF* mutations are frequent oncogenic events in CRC tumors. The mutations lead to constitutively activated proteins that promote cell proliferation by stimulating the RAS/RAF/MEK/ERK kinase cascade (Downward, 2003; Rajagopalan et al., 2003). Furthermore, alterations in the PI3-kinase (PI3K) pathway are also involved in CRC. The PI3K signaling pathway plays a key role in cancer proliferation, survival, motility, and metabolism. Mutations in *PI3 kinase catalytic subunit alpha* (*PIK3CA*) have been reported in approximately 32% of CRCs (Samuels & Velculescu, 2004; Samuels & Ericson, 2006). Ligand binding to epidermal growth factor receptor (EGFR) leads to the activation of both MAPK and PI3K pathway signaling. Treatment with cetuximab, a monoclonal antibody directed against EGFR, is routinely used in the clinic to improve patient survival. However, patients who have *KRAS* mutations are resistant to anti-EGFR antibody treatments (Misale et al., 2012).

The loss of TP53-mediated pathways of apoptosis is another important event in the progression from adenoma to malignant tumor. The premalignant cell attempts to block the development of cancer through the functions of the TP53 protein. TP53 can cause cells to enter quiescence or apoptosis in the event that the machinery regulating cell proliferation is disturbed or the cell is exposed to different types of physiological stress. Mutations in *TP53* are thought to occur relatively late in the development of colorectal tumors. Elimination of TP53 functions is greatly beneficial for the cancer cells and allows the cells to liberate themselves from its cytostatic and pro-apoptotic effects (Baker et al., 1990). In MSI CRCs, *BAX* is frequently inactivated (Rampino et al., 1997; Ionov et al., 2000). The pro-apoptotic protein BAX, which belongs to the bcl-2 family, functions as a key effector of the mitochondrial apoptotic pathways and participates in executing TP53-mediated apoptosis (Chipuk et al., 2004; Miyashita & Reed, 1995).

Alteration of the TGF-β signaling pathway is critical for CRC progression and regulates epithelial proliferation, differentiation, invasion and apoptosis. The signaling cascade is initiated when TGF-β ligands bind to a heterodimeric receptor, consisting of serine/threonine receptors I and II, on the cell surface. This activates the type I receptor kinase activity, which leads to phosphorylation of receptor-specific SMADs, such as SMAD2. Phosphorylated SMADs are then translocated to the nucleus by co-SMADs, such as SMAD4. In the nucleus, the SMAD complex, in concert with other DNA-binding proteins, activates the transcription of specific target genes, for example *P21* and *JUNB* (Massagué, 2008). TGF-β signaling pathway has a key role in controlling normal colonic epithelial homeostasis, and gene

alterations in this pathway are found both in sporadic and hereditary forms of CRC (Xu & Pasche, 2007).

3.3 Inherited predisposition to colorectal cancer

A comprehensive analysis on twins estimated that inherited factors contribute 35 % to the risk of developing CRC. The relative risk in siblings of patients affected by CRC is 2-3 fold (Lichtenstein et al., 2000). A large proportion of the inherited susceptibility to CRC is still unexplained. The distribution of allelic effects involved in complex traits, such as CRC, is predicted to be L-shaped. That is, a small number of alleles confer a large effect on the phenotype, whereas the great majority of alleles have individually a small effect. Rare risk alleles with large effects predominantly predispose individuals to hereditary cancer syndromes or more extreme phenotypes, e.g. very young age of onset (Bost et al., 2001; Mackay, 2001).

3.3.1 Hereditary colorectal cancer syndromes

Hereditary CRC syndromes (Table 2) are thought to explain less than 5 % of all CRC cases in the population (Aaltonen et al., 2007). CRC syndromes have classically been divided into two groups based on the presence or absence of gastrointestinal polyps. The polyposis syndromes, usually identified clinically, are defined by the presence of multiple polyps in the colon. The most common CRC syndromes are Lynch syndrome (also known as hereditary non-polyposis colorectal cancer, HNPCC) (Aaltonen et al., 1998; Lynch et al., 2006) and familial adenomatous polyposis (FAP)(Bodmer et al., 1987)

3.3.1.1 Lynch syndrome

Lynch syndrome, also referred to as hereditary non-polyposis colorectal cancer (HNPCC), is a dominantly inherited syndrome that accounts for approximately 2-5 % of all CRC cases (Aaltonen et al., 1998; Lynch et al., 2006). In Lynch syndrome families multiple generations are usually affected with CRC at an early age (around 45 years). Lynch syndrome patients develop tumors predominantly in the proximal colon and show an excess of synchronous and metachronous cancers. In addition, they have an increased risk for extracolonic cancers, such as endometrial, ovarian, gastric, and pancreatic cancers (Lynch, 1999; Aarnio et al., 1999). The penetrance of this syndrome is high; the lifetime risk of developing CRC is up to 80% in men and 50% in women. In addition, women have approximately a 40-60% risk of developing endometrial cancer (Aarnio et al., 1995; Stoffel et al., 2009).

Table 2. Summary of hereditary CRC syndromes (modified from Kilpivaara and Aaltonen 2013)

Syndrome	Gene(s)	Gene Function
Lynch Syndrome	*MLH1* (Lindblom et al., 1993; Bronner et al., 1994; Papadopoulos et al., 1994)	DNA mismatch repair
	MSH2 (Peltomäki et al., 1993, Fishel., 1993; Leach et al., 1993)	DNA mismatch repair
	MSH6 (Miyaki et al., 1997)	DNA mismatch repair
	PMS2 (Nicolaides et al., 1994)	DNA mismatch repair
Familial adenomatous polyposis	*APC* Bodmer et al., 1987; Kinzler et al., 1991; Nishisho et al., 1991; Groden et al., 1991; Joslyn et al., 1991)	Wnt signaling
Peutz-Jeghers syndrome	*LKB1/STK11* (Hemminki et al., 1997; Amos et al., 1997; Hemminki et al., 1998)	Activation of AMPK-related kinases
Juvenile polyposis	*SMAD4* (Howe et al., 1998a; Howe et al., 1998b)	Signal trunsduction of the TGF-β superfamily and BMPs
	BMPR1A (Howe et al., 2001)	TGF-β signaling
MYH-associated polyposis	*MUTYH* (Al-Tassan et al., 2002; Sieber et al., 2003)	DNA base excision repair
Colorectal cancer and familial tooth agenesis	*AXIN2* (Lammi et al., 2004)	Wnt signaling
Polymerase proofreading-associated polyposis	*POLD1* (Palles et al., 2013)	Catalytic and proofreading subunit of DNA polymerase δ1
	POLE (Palles et al., 2013)	Catalytic subunit of DNA polymerase ε

In 1993, the first susceptibility locus for this syndrome was mapped to 2p16 by using linkage analysis (Peltomäki et al., 1993) and the *MSH2* predisposing gene was subsequently identified (Fishel et al., 1993; Leach et al., 1993). At the same time, Aaltonen *et al*. (1993) reported microsatellite instability in the tumors of the patients, linking Lynch syndrome to defective mismatch repair. Later, *MLH1* was mapped to 3p21 (Lindblom et al., 1993; Bronner et al., 1994; Papadopoulos et al., 1994), and the gene was subsequently shown to be the most important Lynch syndrome gene since it accounts for half of all cases (Peltomäki & Vasen, 2004). In Finland, there are two founder mutations in *MLH1* that explain up to 60% of all Lynch syndrome cases (Nyström-Lahti et al., 1995; 1996). Other genes that have been linked to Lynch syndrome predisposition are *MSH6* (Nicolaides et al., 1994) and *PMS2* (Miyaki et al., 1997). Patients carrying germline *PMS2* mutations show clinical features of Turcot syndrome, characterized by familial aggregation of primary brain tumors in addition

to colorectal adenomas and carcinomas (Peltomäki, 2005).

All the identified genes encode key components of the mismatch repair system, which consists of two main heterodimeric protein complexes: the MutL homologue (MLH1 and PMS2) and the MutS homologue (MSH2 and MSH6) (Kolodner et al., 1995). A two-hit mechanism for DNA mismatch repair (MMR) gene inactivation has been shown to underlie the microsatellite instability (MSI) phenotype. Tumors from Lynch syndrome patients frequently show silencing of the MMR gene through LOH involving the wild-type allele (Parsons et al., 1993; Hemminki et al., 1994).

Table 3. Diagnostic criteria for Lynch Syndrome (HNPCC) and guidelines for MSI testing

Amsterdam criteria II (Revised ICG-HNPCC Criteria)*
1) There should be at least three relatives with an HNPCC-associated cancer (CRC, endometrium, small bowel, ureter, or renal pelvis) 2) One 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 before 50 years of age 5) FAP should be excluded in all CRC cases
Bethesda guidelines (Revised)**
Tumors from individuals should be tested for MSI in the following situations: 1) CRC diagnosed in a patient who is less than 50 years of age 2) Presence of synchronous, metachronous colorectal, or other HNPCC-associated tumors, regardless of age 3) CRC with MSI histology (tumor inflitrating lymphocytes, Crohn's-like lymphocytic reaction, mucinous/signet-ring differentiation or medullary growth pattern) diagnosed in a patient who is less than 60 4) CRC diagnosed in ≥ two first- or second degree relatives with HNPCC-related tumors, regardless of age

* Vasen et al., 1999, ** Umar et al., 2004

Several international diagnostic criteria have been developed for Lynch syndrome, the foremost being Amsterdam I (Vasen et al., 1991), Amsterdam II (including extracolonic cancers) (Vasen et al., 1999) and Bethesda Guidelines (Rodrigues et al., 1990; Umar et al., 2004) (Table 3). Since Lynch syndrome patients do not display any distinct clinical features, definite diagnosis can only be done by demonstrating the presence of a germline MMR mutation. Initial screening is often performed by analyzing MSI status in the tumors, combined with immunohistochemical staining (Lynch & Lynch, 2005). Bethesda guidelines were developed to identify individuals who should be tested for MSI. The Bethesda panel used for PCR-based MSI testing consists of five microsatellite markers; two out of five markers displaying length alterations is classified as MSI high (Rodrigues et al., 1990; Umar et al., 2004). Lynch syndrome patients should regularly be screened for colorectal adenomas and carcinomas by colonoscopy. It is recommended that patients undergo colonoscopy every 1-3 years starting from the age of 20-25 and any observed adenomas should be

removed (Mecklin & Järvinen, 2005).

3.3.1.2 Familial adenomatous polyposis

Familial adenomatous polyposis (FAP) is an autosomal dominant disorder accounting for 0.5-1% of all CRC cases (Burn et al., 1991; Järvinen, 1992). The patients have very distinct clinical features, with hundreds of thousands of adenomas throughout the colon, and, if not treated correctly, will inevitably develop CRC at an early age (35-40 years). FAP patients have also an increased risk for other malignancies, including pigmented lesions of the retina, osteomas of the jaws, desmoid tumors and papillary thyroid carcinoma (Galiatsatos & Foulkes, 2006). Attenuated FAP is a subtype of classical FAP where patients display fewer colonic adenomas and usually at a later age (Lynch et al., 1995).

FAP is caused by germline mutations in the *APC* gene on 5q21 (Bodmer et al., 1987; Kinzler et al., 1991; Nishisho et al., 1991; Groden et al., 1991, Joslyn et al., 1991). Many mutations have been identified, mostly nonsense and frameshift mutations, and there is a correlation between the genetic site of the mutations and the severity of the clinical manifestations (Rozen et al., 1999; Friedl et al., 1996). Also, reports have shown that the type of the second hit depends on the random first hit in the germline. It is believed that this occurs in order for the cell to ensure optimal WNT signaling levels (Lamlum et al., 1999; Albuquerque et al., 2002). Some cancers have been shown to harbor a third hit at *APC*, mostly through copy number gains and losses, in order to further fine tune the level of WNT activation (Segditsas et al., 2009).

FAP can be diagnosed clinically by the demonstration of > 100 colorectal adenomas. Genetic testing should be performed on all individuals showing classical polyposis to confirm the diagnosis. Also individuals with a family history of CRC or those with fewer adenomas, which might exhibit attenuated forms of the disease, should undergo genetic screening. Annual endoscopy is suggested for all mutation carries. Colectomy can be considered as a prophylactic treatment option (Järvinen, 2003).

3.3.1.3 Other syndromes

Other rare autosomal dominant CRC syndromes are Peutz-Jeghers syndrome (PJS) and Juvenile Polyposis (JP) that are both characterized by intestinal hamartomatous polyposis (Schreibman et al., 2005). PJS is caused by mutations in *LKB1,* encoding a serine/threonine kinase, located on chromosome 19q (Hemminki et al., 1997; Amos et al., 1997; Hemminki et al., 1998). Polyps of PJS patients are mostly found in the small intestine and the stomach, but also occur in the colon (Hemminki et al., 1998). Germline mutations in *SMAD4* on 18q (Howe et al., 1998a; Howe et al., 1998b) and *BMPR1A* on 10q (Howe et al., 2001), members of the TGFβ-superfamily, underlie JP.

Polyps of JP patients are most prevalent in the colon and rectum; however, they may also affect the gastrointestinal tract. The age of onset is usually at teenage years (Howe et al., 1998).

The MYH-associated polyposis syndrome is the only known CRC syndrome that is inherited in a recessive manner. Patients have a clinical phenotype similar to that of attenuated FAP patients; they display 10-100 colonic adenomas. The age of onset is typically around 46 years (Jo & Chung, 2005). The disease is caused by biallelic mutations in the base-excision repair gene *MYH* (also known as *MUTYH*) (Al-Tassan et al., 2002; Sieber et al., 2003). Also, monoallelic carriers have been reported to have a slight increase in CRC risk (Jenkins et al., 2006).

Recently, a whole-genome effort discovered germline mutations in *POLE* and *POLD1,* which encode subunits of polymerase ε and δ, to underlie a rare novel syndrome termed polymerase proofreading–associated polyposis. Individuals with these mutations present multiple or large adenomas and adenocarcinomas (Palles et al., 2013). Heterozygous germline mutations in *AXIN2* have been associated with a syndrome characterized by familial tooth-agenesis and predisposition to CRC (Lammi et al., 2004).

3.3.2 Low- and moderate-penetrance alleles

Inherited factors are estimated to play a role in approximately one-third of all CRC cases (Lichtenstein et al., 2000). High-penetrance mutations in known CRC predisposing genes explain only a small fraction of all cases with inherited susceptibility. Still today, the great majority of common CRC families, with only one first-degree relative affected, remain molecularly unexplained (Aaltonen et al., 2007). In Finland, around 11% of all patients with CRC have at least one first-degree relative with CRC (Salovaara et al., 2000). It is estimated that more than 60% of the excess familial risk remains molecularly unexplained (Lubbe et al., 2009; Salovaara et al., 2000). Genetic susceptibility to CRC underlies an unknown proportion of both familial and sporadic cases, and the division of cases into one of these two groups might be somewhat arbitrary. A few, hundreds or even thousands of predisposition alleles with different levels of risk and prevalence in the population are likely to collectively contribute to CRC susceptibility, accounting for both familial CRC cases as well as sporadic cases. In addition, modifier genes are also likely to influence the effect of genetic and environmental factors that contribute to CRC (la Chapelle, 2004).

Over the past 15 years, several strategies have been applied to identify additional predisposing variants. To date, several GWA studies on CRC have been performed. In total, common low-penetrance variants at approximately 20 genomic loci have been identified to be associated with CRC susceptibility. However, the loci detected

by GWA studies confer only a very modest effect on CRC risk, typically with odds ratios < 1.5 (Dunlop et al., 2012; 2013; Lubbe et al., 2012). For the great majority of loci the causative variants remain to be identified. One proposed mechanism of action is that the common variants would influence distal enhancer elements that regulate expression of key target genes (Pomerantz et al., 2009; Maurano et al., 2012). 8q24, with the polymorphism rs6983267, is one of the most interesting CRC regions pinpointed by GWA studies (Tomlinson et al., 2007). At this region, a cancer-specific enhancer element has been identified that has been suggested to control the expression of the *MYC* oncogene (Tuupanen et al., 2009; Sur et al., 2012). Although, the effect of each common low-penetrance variant on CRC risk is small, an additive contribution has been observed. Studies have estimated that ten known low-penetrance variants collectively explain less than 9% of the variance in familial CRC risk (Houlston et al., 2008; Niittymäki et al., 2010).

The fact that common variants only explain a very small part of the variance in risk has led to alternative views of where to find the "missing heritability". It has been proposed that rare variants of larger effects or common variants of very small effects might explain most of the variance in risk (Fletcher & Houlston, 2010; Gibson, 2011). Despite numerous candidate gene-screens, very few rare or low frequency variants of moderate penetrance have been identified for CRC. The majority of the proposed associations have not reached statistical significance and have been restricted by small sample sizes. Perhaps the only well-established example of a moderate-penetrance variant is the *APC* I1307K that is carried by approximately 6 % of Ashkenazi Jews and confers a two-fold increase in CRC risk (Laken et al., 1997).

Common CRC families, with only few affected cases, form an attractive patient group to search for additional predisposition variants. However, these families are usually too small for traditional linkage analysis and the variants are likely to be too diverse and rare to be detected by GWA studies. Advances in sequencing technologies have made exome and whole genome sequencing attractive approaches for identification of rare variants of varying penetrance. Recently, two small-scale exome sequencing efforts were conducted to search for novel CRC predisposition variants. In the limited sequence data, no variants were significantly associated with CRC. Both studies discovered a small number of genes that remain good candidates for CRC predisposition. No gene was identified as a candidate in both studies (Derycke et al., 2013; Smith et al., 2013). However, validation in larger sample sets and in other populations will be required for providing evidence for their association with CRC.

AIMS OF THE STUDY

The main aim of this study was to provide a better understanding of the molecular mechanisms behind CRC predisposition and progression. The specific aims are listed below.

Somatic mutation in MSI CRC (I-II)

I. To characterize nonsense-mediated decay-escaping target genes with potentially oncogenic effects

II. To identify novel oncogenes with mutation hot spots by exome sequencing

Germline variants in familial CRC (III-IV)

III. To study the role of 15 candidate cancer genes in familial CRC predisposition

IV. To identify novel CRC susceptibility genes for familial CRC by utilizing exome sequencing

MATERIALS AND METHODS

The materials and methods used in this study are presented shortly below and are described in more detail in the original publications (referred to here by their roman numbers).

1 Sample material

Colorectal cancer patients (I-IV). A population-based material of 1,042 CRC patients was collected between 1994 and 1998 from nine Finnish central hospitals (Aaltonen et al., 1998; Salovaara et al., 2000). After 1998, sample collection was continued from two Finnish central hospitals and material from 472 CRC patients were available from this series (unpublished collection). The materials included normal and tumor tissue. Tumor samples have been studied for MSI, and in positive cases *MLH1* and *MSH2* have been Sanger sequenced. Detailed clinical information and pathological evaluations were available for all cases. Data on all first-degree relatives and their cancer occurrence were acquired from the Finnish Population Registry (http://www.vrk.fi) and the Finnish Cancer Registry (http://www.cancer.fi/syoparekisteri/en/). Signed informed consent or authorization from the National Supervisory Authority for Welfare and Health was obtained for all samples. These efforts were reviewed and approved by the Ethics Committee and the Hospital district of Helsinki and Uusimaa (HUS).

In study I, 100 MSI CRCs were selected for mutation screening. The entire coding regions of highly mutated genes were screened in a set of 30 MSS CRCs. For clinicopathological associations, an additional set of 31 MSI CRC DNAs extracted from paraffin-embedded tumors obtained from Dr. Markus Mäkinen (Oulu University Hospital) were used. In study II, the exomes of 25 sporadic MSI CRC-normal pairs were sequenced as a discovery set. For validation, a sample set of 86 MSI tumors were available. Top-ranked genes were also screened in 75 MSS CRCs. In study III and IV, mutation screening was performed in familial CRC cases. All cases fulfilled the following criteria: 1) at least one CRC case in first degree relatives, 2) negative for any known high-penetrance CRC gene and 3) availability of sufficient amount of DNA extracted from both tumor and normal tissue for respective studies. In study III, candidate gene screens were performed in 45 familial cases. Additional genotyping was performed in 967 population-matched CRC cases. In study IV, the discovery set included 96 independent familial CRC cases. The validation phase samples consisted of 954 population-matched cases.

Population matched controls (III-IV). DNA samples from population matched anonymous blood donors were used as controls. These were obtained from the Finnish Red Cross Blood Transfusion Service.

Cell lines (I-II). MSI CRC cell lines DLD1, GP5D, HCA7, HCT116, HCT15, HCT8, HUTU80, LIM1215, L174T, RKO, SNUC2B, VACO5, CCL-231, LoVo, LS180 and human embryonic kidney cell line HEK293 were obtained from the American Type Culture Collection (AACR), the European Collection of Cell Culture (ECACC), or provided by Professor Ian Tomlinson. Commercially available Flp-In 293 T-Rex cells (Invitrogen) were also used in study II.

2 Genetic analyses

DNA (I-IV) and RNA (I-III) extractions. Genomic DNA from peripheral blood or fresh frozen tissue samples was extracted by a previously described non-enzymatic method (Lahiri & Nurnberger, 1991). Total RNA was extracted with the RNeasy Kit (Qiagen) or with TRIzol reagent (I-III; Invitrogen). cDNA synthesis was performed with M-MLV enzyme (Promega).

In study I, laser-capture microdissection on malignant epithelial cells was performed prior to DNA extraction. Fresh-frozen tumor sections were prepared and stained with HistoGene LCM Frozen Section Staining Kit (Arcturus). Laser capture microdissection was performed on malignant epithelial cells with Arcturus Microdissection Instrument on CapSure LCM Caps (Arcturus). Genomic DNA was extracted by PicoPure DNA Extraction Kit (Arcturus).

PCR and Sanger sequencing (I-IV). Sequencing primers were designed with Primer3 (http://biotools.umassmed.edu/bioapps/primer3_www.cgi) and Primer3 Plus programs (http://www.bioinformatics.nl/cgi-bin/primer3plus/primer3plus.cgi/) and reference sequences were obtained from the Ensembl (http://www.ensembl.org) and NCBI (http://www.ncbi.nlm.nih.gov) databases. Fragments were amplified with AmpliTaqGold enzyme (Applied Biosystems) or Phusion DNA polymerase (Finnzymes). PCR products were purified by using the ExoSAP-IT PCR purification kit (USB Corporation). Sanger sequencing was performed using BigDye v3.1 sequencing chemistry (Applied Biosystems) and 3730xl DNA Analyzer (Applied Biosystems).

Genotyping (III, IV). Genotyping was carried out by using the 7900HT Fast Real-Time PCR System (Applied Biosystems) and massARRAY iPLEX Gold (Sequenom) and performed at the Institute for Molecular Medicine Finland (FIMM, University of Helsinki) and Estonian Genome Center (University of Tartu).

Exome sequencing (II, IV). Coding regions were enriched with the Agilent SureSelect Human All Exon Kit v1 (Agilent) according to the manufacturer's instructions. Sequencing of paired end short reads was performed on either Illumina

GAII or HiSeq platforms (Illumina) at Karolinska Institutet (Sweden) and at Finnish Institute for Molecular Medicine (FIMM) Genome and Technology Center. Raw sequencing data was run through an in-house analysis pipeline for exome sequencing data, which consisted of the following programs and tools: FASTQC (http://www.bioinformatics.bbsrc.ac.uk/projects/fastqc), Burrow-Wheelers Aligner (Li & Durbin, 2009), Picard Tools Markduplicates (http://picard.sourceforge.net), Samtools mpileup (Li et al., 2009), Genome Analysis Toolkit IndelRealigner (McKenna et al., 2010) and Genome Analysis Toolkit UnifiedGenotyper (McKenna et al., 2010). In addition, an in-house script was utilized to remove 3' ends with high adapter similarity. Exome data was analyzed with a visualization and comparative analysis tool developed in the laboratory (unpublished) as described below.

In study II, somatic mutations from 25 MSI CRC exomes were extracted by filtering the data against exome data from the respective normal tissues. Somatic sequence data was control filtered against data from the 1000 Genomes Project (Phase 1 release (1000 Genomes Project Consortium et al., 2012)), population matched exome control data (n=69) and data from the Database of Single Nucleotide Polymorphisms (Build 132, www.ncbi.nlm.nih.gov/SNP) to further exclude germline variants. Subsequent analysis focused on non-synonymous missense variants forming potential mutation hot spots. Sanger sequencing was performed on all missense variants that located to the same or adjacent codons in at least two tumors.

In study IV, the germline exome data from 96 familial CRC cases was analyzed for putative loss-of-function variants, including nonsense, frameshifting insertions and deletions, and splice-site variants. Variants were filtered against population matched exome control data (n=212) and data from the 1000 Genomes Project (Phase 1 release, (1000 Genomes Project Consortium et al., 2012)) with a MAF threshold of 0.001. Subsequent analysis was focused on genes with putative loss-of-function variants in at least two cases. These were further validated by Sanger sequencing.

Expression microarray analysis (I). A previously characterized Human Genome U133A 2.0 oligonucleotide microarray (Affymetrix) dataset was available for this study (Andersen et al., 2009). RNA samples from 73 MSI CRCs and 10 normal colonic mucosa samples were analyzed. Analysis was limited to genes that were overexpressed in MSI CRCs compared to normal mucosa (mean fold change 2).

Quantitative real-time PCR (II). Relative expression of mRNA was determined with TaqMan chemistry and the ABI Prism 7500 sequence detection system (Applied Biosystems). Assays for *ZBTB2* (assay ID, Hs00535603_m1), *PSRC1* (Hs00364137_m1), *RANBP2* (Hs01108576), and *Human B-actin* (4310881E) were used (Applied Biosystems).

3 Protein analyses

Western blot analysis (I, II). Proteins (25-30*ug*) were fractioned by sodium dodecyl sulfate-polyacrylamide gel electrophoresis in a 7.5% Tris-HCL polyacrylamide gel and transferred to a nitrocellulose filter. Immunoblots were incubated with primary antibodies: anti-mouse TTK (H00007272-M01; Abnova) and anti-human influenza hemagglutinin (HA-11; Covance) at dilutions 1:500 and 1:10,000, respectively. Antibodies against alpha-tubulin (T5168; Sigma-Aldrich) and glyceraldehyde-3-phosphate dehydrogenase (ab9485; Abcam) were used as loading controls. Appropriate horseradish peroxidase conjugated secondary antibodies were used followed by enhanced chemiluminescence detection.

Immunohistochemistry (I). Paraffin-embedded tumor sections were stained with a monoclonal antibody against human TTK (Zymed Laboratories Invitrogen) for 2 hours at a dilution of 1:50. Tris-EDTA buffer (pH 9) was used in the heat-induced epitope retrieval. Detection was performed with anti-Mouse/Rabbit PowerVision Poly-HRP IHC Detection System (Leica Biosystems Newcastle Ltd).

Immunofluorescence (II). Fixed cells were stained with the anti- HA antibody (Covance) and secondary Alexa Fluor-488-conjugated goat anti-mouse IgGs (Invitrogen, Life Technologies) at dilutions 1:10,000 and 1:500, respectively. Nuclei were also stained with Hoechst stain solution (Sigma-Aldrich). The Zeiss Axioplan 2 upright epifluorescence microsope was used for imaging.

Liquid chromatography-mass spectrometry (II). Large-scale protein-complex analysis was performed on ZBTB2 and PSRC1, both wild-type and mutant forms. Both sample preparations and liquid chromatography-mass spectrometry runs were performed in the Institute of Biotechnology (Finland) in collaboration with docent Markku Varjosalo. Single-step affinity purification was performed as described previously (Varjosalo et al., 2013). In short, for each pulldown, a cell pellet deriving from 5 × 15–cm fully confluent dish was lysed for 10 min on ice in 5 mL HNN lysis buffer. Replicates were done for each pulldown. Bait proteins were tagged with a Twin-Strep-tag (“SH”), and strep-Tactin Sepharose beads were used to isolate the bait proteins. The beads were transferred to a Bio-Spin chromatography column (Bio-Rad) and washed with 3 × 1 mL HNN buffer and 3 × 1 mL HNN buffer without detergent and inhibitors, and bound proteins were natively eluted into fresh Eppendorf tubes. Mass spectrometry runs were performed on an Orbitrap Elite ETD mass spectrometer (Thermo Scientific) with Thermo Scientific nLCII nanoflow system (Thermo Scientific), as described in Varjosalo et al. (2013). Proteome Discoverer software (Thermo Scientific) was utilized both in peak extractions and subsequent protein identification. The data analysis program SEQUEST was used to search for calibrated peak files against databases of human protein sequences

(http://fields.scripps.edu/sequest/).

4 Cell culture studies

Mutagenesis and cloning (I, II). In study I, wild-type and mutant *TTK* cDNA were cloned into pEGFP-C2 vector (Clontech) with an N-terminal green fluorescence protein (GFP) tag. Plasmids were transfected into HEK293 cells with Lipofectamine2000 (Invitrogen) according to the manufacturer's instructions. For localization experiments, paclitaxel (Molecular Probes) was added 24h after transfection (20nM) and cells were subsequently fixed after 6h.

In study II, site-directed mutagenesis was performed according to manufacturer's protocols with QuikChange Site-Directed Mutagenesis Kit (Agilent Technologies) on cDNA clones (*PSRC1*; RG221688, *ZBTB2*; RG204198, OriGene) to generate mutants. The Gateway cloning system (Invitrogen, Life Technologies) was used to clone wild-type and mutant cDNAs into pDONR21 vector (Invitrogen, Life Technologies). This was done according to the manufacturer's instructions. Expression vectors for tetracycline-controlled expression of streptavidin-binding peptide hemagglutinin (HA)-tagged version of the cDNA were produced. An LR recombination was done between the entry clones and the destination vectors: pcDNA5/FRT/TO/SH/GW (N-terminally tagged) and pcDNA5/FRT/TO/cSH (N-terminally tagged). Destination vectors were designed in the laboratory of Dr. Markku Varjosalo (Varjosalo et al., 2013).

Generation of stable and inducible cell lines (II). Flp-In 293 T-REx cells (Invitrogen) were cultured as described by the manufacturers. Cells were co-transfected with the expression vectors (pcDNA5/FRT/TO/SH/GW and pcDNA5/FRT/TO/cSH) and the pOG44 vector (Invitrogen) using Fugene transfection reagent (Roche). Hygromycin (100 μg/mL) was added two days after transfection for selection. Positive and isogenic clones were collected and amplified after 2-3 weeks. Inducible expression, by incubating the cells with doxycycline (1 μg/mL, Sigma-Aldrich) for 24h, was verified by Western blot analysis and immunofluorescence.

Cell proliferation assay (II). Stable and inducible Flp-In 293 T-REx cells with SH-tagged ZBTB2 and PSRC1 (both wild-type and mutant), were cultured according to the manufacturer's instructions (Invitrogen, Life Technologies). After one week, cells (2.0×10^5) were plated on 6-well plates with media containing 1 μg/mL doxycycline (Sigma-Aldrich). Four replicates were done for each of the construct-containing Flp-In 293 T-REx cell lines. Every three to four days, cells were manually counted and re-plated, for a total of 13 days. The experiment was repeated once.

Paclitaxel sensitivity assays (I). The sensitivity of four MSI CRC cell lines LoVo,

HCT116, DLD1 and HCT8 to paclitaxel was analyzed with IncuCyte live-cell imaging system (Essen Instruments) and colony formation assay. MSI CRC cell lines were cultured on 48-well plates and filmed once per hour with IncuCyte live-cell imaging system (Essen Instruments) for 96h. Paclitaxel was added after 21h (10, 25 and 50nM). Cell confluence and morphology were analyzed from the time-lapse movies. For the colony formation assay cells were grown on 12-well plates and paclitaxel was added (1, 5, 10 and 25 nM) after 48h incubation. Drug-containing media was changed daily and the cells were incubated for eight days in total. Cells were fixed (96% methanol) and stained with 0.05% Crystal Violet (Sigma Aldrich). Imaging and quantifications of colonies were done with GeneGenius bio imaging system, GeneSnap and GeneTools softwares (Syngene). In both assays, DMSO was used as negative control.

5 Statistical analyses and computational tools

***In silico* variant effect predictions (II, III).** The following programs were used to predict the functional consequence of the identified germline variants and somatic mutations: the Ensembl Variant Effect Predictor (McLaren et al., 2010), SIFT (http://sift.jcvi.org/), Polyphen (http://genetics.bwh.harvard.edu/pph/index.html), PolyPhen2 (http://genetics.bwh.harvard.edu/pph2/), NetGene2 (http://www.cbs.dtu.dk/services/NetGene2/) and BDGP Splice Site Prediction tool (http://www.fruitfly.org/seq_tools/splice.html). Multiple DNA alignments were generated with Multiz alignments (Blanchette et al., 2004). EMBOSS Transeq was used to translate genomic sequences into corresponding peptide sequences and the protein sequences were aligned with T-Coffee (http://www.ebi.ac.uk/tools/).

Computational tool for microsatellite-containing genes (I). A computational tool was created that predicted NMD-escape whenever a frameshift-induced termination codon located 55 bps upstream of the last exon-exon junction or in the last exon (Nagy & Maquat, 1998). All transcripts with mutated mononucleotide repeats (6-10 bp) that were predicted to escape NMD were identified from Ensembl (Homo sapiens 45_36g; www.ensembl.org). Further analysis focused on overexpressed genes in MSI CRC.

Protein interactome analysis (II). The mass spectrometry data was analyzed with Ingenuity Pathways Analyses (www.ingen.com) software. Identified proteins were mapped into relevant groups for molecular and cellular function based on Ingenuity's knowledge base. Heat maps were performed with R.

Statistics (I, II, IV). Chi-squared test was utilized to calculate differences between groups of patients or samples. Bonferroni's correction for multiple testing was used to obtain *P*-values for clinicopathological associations (I). In study II, a randomized test of goodness-of-fit was used to test whether the number of mutation hot spots

observed in the exome data differed from that expected. The mutations observed in the exome data were redistributed randomly in the exome according to previously reported nucleotide frequencies and the mutation hot spots were counted (Greenman et al., 2007). The process was repeated and a null distribution was obtained. *P* value was determined by comparing observed counts to the null distribution. The same was done for the hot spot mutation frequencies observed in the validation set. The –log of *P*-values for molecular and cellular functions based on Ingenuity's knowledge base were calculated by Fisher exact test. In study IV, exact binominal test was used to calculate *P*-values for wild-type allele losses.

RESULTS

1 *TTK* is frequently mutated in microsatellite-unstable colorectal cancer

1.1 Identification of *TTK* mutations

We combined microarray expression profiling of MSI CRCs with a bioinformatics search and identified 330 overexpressed genes that were predicted to escape NMD after a deletion in a microsatellite repeat. Sanger sequencing was done in 30 MSI CRCs for repeats in 258 genes. The great majority of the repeats showed low somatic mutation frequency (≤7%). Four genes were mutated in >20% of the 100 MSI CRCs analyzed: *TTK* (*TTK protein kinase*, 67%), *TMEM97* (*Transmembrane protein 97*, 39%), *ARS2* (*Arsenite-resistance protein 2*, 33%), *LENG8* (*Leukocyte receptor cluster member 8*, 31%). Only *TTK* had a mutation frequency higher than expected by chance when compared to non-coding identical control repeats and was thus studied further.

1.2 *TTK* mutation spectra in colorectal cancer

The last exon (22) of *TTK* has a complex repeat that consists of A9-G4-A7 repeats. The mutation frequency was 59% at this locus in a total set of 179 MSI CRCs analyzed (105/179). The complex repeat (A9-G4-A7) was anticipated to be more unstable than a normal mononucleotide repeat alone, thus, the genome was searched for identical non-coding complex repeats to assess the background mutation frequency at such loci. *TTK* was shown to harbor mutations significantly more often than compared to identical control repeats (105/179 in *TTK* vs. 50/139 in controls, P=9.5 x 10^{-5}).

Table 4.*TTK* frameshift mutations in MSI CRC

Exon (repeat type)	MSI CRCs	Mutation (coding)
2 (A5)	1/100	c.86delA
5 (A7, A7)	21/279 (12%)	c.484delA, c575delA
20 (A4)	1/100	c.2331delA
22 (A9-G4-A7)	105/179 (59%)	c.2560delA*

*The most frequent mutation at the repeat

Additional TTK frameshift mutations were observed at repeats in exons 2, 5 and 20 (Table 4). All mutations identified were heterozygous. Exon 5 deletions were enriched in exon 22 wild-type tumors (P=0.017). Exons 5 and 22 were also analyzed for mutations in 12 MSI CRC cell lines and repeat mutations were found at frequencies 25% and 58%, respectively (see Table II in the original publication). Mutation hot spots in exons 5 and 22 were also screened in 848 MSS CRCs and three

tumors showed frameshift mutations in exon 22.

1.3 Expression and localization of TTK

Western blot analysis showed the most common *TTK* frameshift mutation (c.2560delA) to result in elongation of the peptide by 34 amino acids in two heterozygous mutant cell lines, LoVo and HCT116 (see Figure 4A). Immunohistochemical stainings of TTK revealed no difference in localization or intensity between wild-type and mutant MSI CRC tumors (see Figure 1D in original publication). GFP fusions of the wild-type and mutant (c.2560delA) TTK proteins were expressed in HEK293 cells and showed similar localization; TTK localized at the cytoplasm and at the kinetochores (see Figure 3 in the original publication).

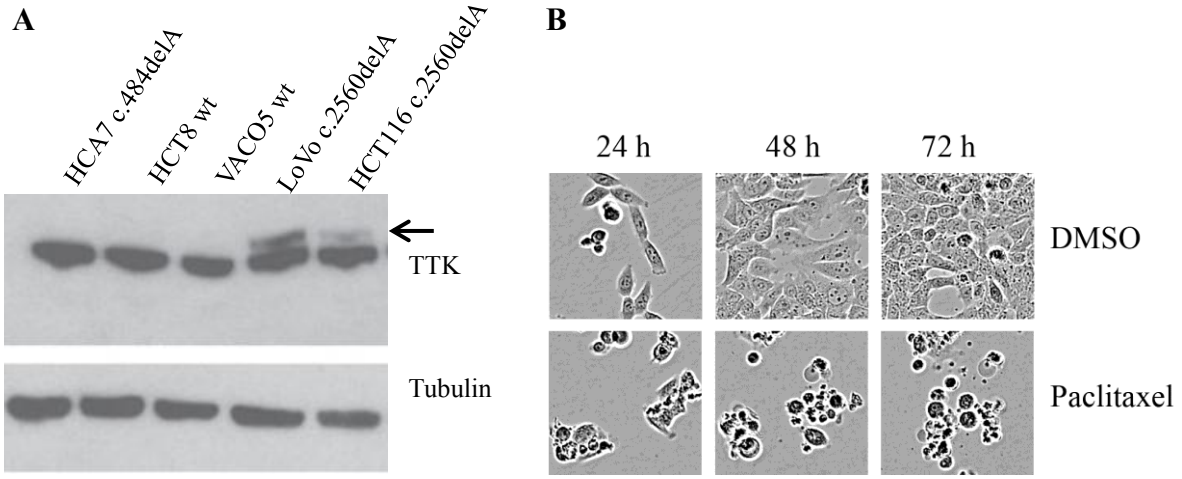

Figure 4. Functional studies on mutant TTK. (**A**) The elongated TTK peptide (arrow) was detected in LoVo and HCT116 MSI cell lines with the c.2560delA mutation. (**B**) The mutant cell line HCT116 was incubated with the microtubule-stabilizing drug paclitaxel (10nM) and DMSO, as control, to test the spindle-assembly checkpoint arrest. Examples of live-cell images shown from time-points 24h, 48h and 72h. Cells cultured with paclitaxel arrested in the checkpoint and typically died from this arrest.

1.4 TTK and the spindle assembly checkpoint

Two wild-type (DLD1 and HCT8) and two mutant MSI CRC cell lines (LoVo and HCT116) were treated with the microtubule-stabilizing drug paclitaxel and analyzed with IncuCyte live-cell imaging and colony formation assay. When cells were cultured in the presence of paclitaxel, a reduction in confluence was observed by live-cell imaging in the mutant cells. Both LoVo (figure not shown) and HCT116 cells (Figure 4B) begun to arrest in mitosis and later mostly died from the arrest, suggesting that the checkpoint response is intact in mutant MSI CRC cells. In the colony formation assay, viability was drastically reduced for HCT116 when treated with paclitaxel, providing further evidence that the checkpoint response is

functional (see Figure 4 in the original publication). In the LoVo cells, colony formation was insufficient for quantification even in the control and was therefore excluded.

2 Novel candidate oncogenes in microsatellite-unstable colorectal cancer

2.1 Identification of fifteen candidate oncogenes with mutation hot spots

We systematically searched for novel oncogenes by exome sequencing 25 MSI CRCs and respective healthy tissues. All cases were sporadic. The exome data was searched for potential hot spot mutations, in other words, missense mutations in at least two tumors hitting the same or adjacent codon. As expected, hot spot mutations in known MSI CRC oncogenes were identified: *BRAF* (V600, 32%, 8/25), *KRAS* (8%, 2/25) and *CTNNB1* (8% 2/25). In addition, novel potential mutation hot spots were observed in 30 genes in the exome data. Mutation hot spots in 15 genes were confirmed by Sanger sequencing (Table 5). In the discovery set of 25 MSI tumors, *ITGA7* displayed three hot spot mutations, and the rest of the genes showed two.

Table 5. Novel candidate oncogenes in MSI CRC

Candidate oncogene	Gene description*	Mutation hot spot site (amino acid)
ADAR	*adenosine deaminase, RNA-specific*	Tyr1173, Arg1172
DCAF12L2	*DDB1 and CUL4 associated factor 12-like 2*	Arg335
GLT1D1	*glycosyltransferase 1 domain containing 1*	Ala157, Val158
ITGA7	*integrin, alpha 7*	Ala970, Arg969
MAP1B	*microtubule-associated protein 1B*	Pro480, Ala481
MRGPRX4	*MAS-related GPR, member X4*	Ser114, Ala115
PSRC1	*proline/serine-rich coiled-coil 1*	Arg136, Thr135
RANBP2	*RAN binding protein 2*	Arg945
RPS6KL1	*ribosomal protein S6 kinase-like 1*	Thr55, Ala54
SNCAIP	*synuclein, alpha interacting protein*	Arg499
TCEAL6	*transcription elongation factor A (SII)-like 6*	Pro101, Arg100
TUBB6	*tubulin, beta 6 class V*	Ala231, Thr232
WBP5	*WW domain binding protein 5*	Arg46, Glu47
VEGFB	*vascular endothelial growth factor B*	Val54, Pro55
ZBTB2	*zinc finger and BTB domain containing 2*	Arg262, Arg261

Genes with mutations in the validation set	Identified Mutations (cDNA)**	Overall frequency of hot spot mutations
PSRC1	c.404G>A (x2), c.407C>T, c.406G>A (x2)	5/109, 4.6%
RANBP2	c.2833C>T (x2), c.2834G>A (x2)	4/109, 3.7%
ZBTB2	c.781G>A (x3), c.784G>A	4/106, 3.8%

* Gene descriptions taken from HGNC (http://www.genenames.org)

** In brackets the number of times the mutation was identified

Interestingly, only three genes showed additional hot spot mutations in the validation set: *ZBTB2*, *RANBP2* and *PSRC1* with mutations frequencies of 2.5% (2/81 MSI CRCs), 2.4% (2/84) and 3.6% (3/84), respectively. The overall frequencies of hot spot mutations were 3.8% (4/106 MSI CRCs), 3.7% (4/109) and 4.6% (5/109) for *ZBTB2*, *RANBP2* and *PSRC1*, respectively (Table 5). In Figure 5, the domains and missense mutation sites of ZBTB2, RANBP2 and PSRC1 proteins are depicted. Mutation sites in *ZBTB2* and *RANBP2* were highly conserved across species (see Figure 1 and Supplementary Figure 1 in the original publication). No hot spot mutations were identified in 75 MSS CRCs and 12 MSI CRC cell lines in these three genes. Overexpression of *ZBTB2*, *RANBP2* and *PSRC1* was observed in mutation-positive tumors (in 2/2, 2/2 and 2/4 tumors, respectively) compared to respective normal tissue (see Supplementary Figure 1 in the original publication).

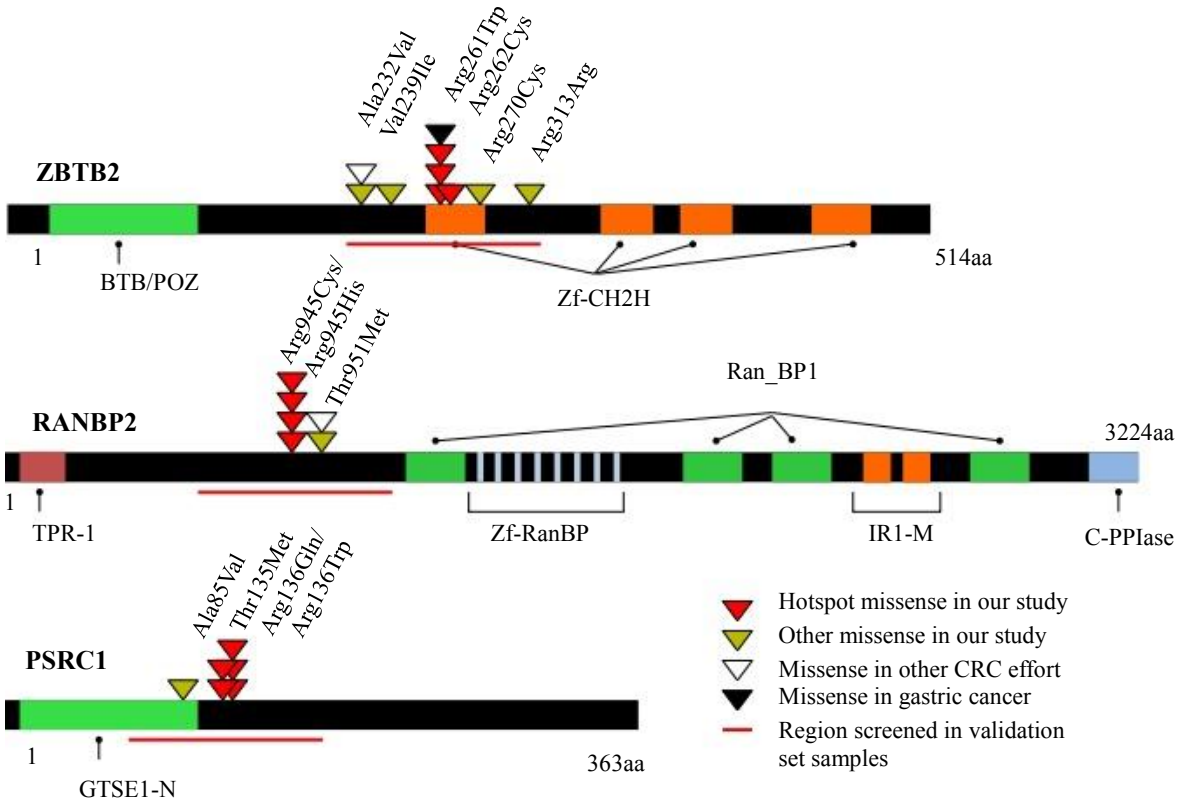

Figure 5. Schematic diagram of the domains and missense mutations of ZBTB2, RANBP2 and PSRC1. The red line indicates the region that was screened in the validation set samples. Data from Seshagiri et al., 2012 and Cancer Genome Atlas Network, 2012 were searched for the same missense mutations identified here (white arrow). Sequencing data from other cancer types in the cBioPortal for Cancer Genomics (http://www.cbioportal.org/public-portal/) were searched for our hot spot mutations (black arrow). Information on structural and functional domains was obtained from cBioPortal for Cancer Genomics (http://www.cbioportal.org/public-portal/) and the Ensembl database (http://www.ensembl.org). BTB: BR-C, ttk and bab; POZ: Pox virus and Zinc finger; Zf-CH2H: Zinc finger C2H2 type; TPR-1: tetratricopeptide repeat domain 1; Zf-RanBP: Zn-finger in Ran binding protein; Ran_BP1: Ran binding domain; IR1-M:internal repeat

domain; C-PPIase: Cyclophilin-like peptidyl-prolyl cis-trans isomerase domain; GTSE1-N: G-2 and S-Phase Expressed 1; Aa: Amino acid.

2.2 Functional studies on mutant ZBTB2 and PSRC1 proteins

Next, we wanted to investigate the impact of the hot spot mutations on protein functions by mass spectrometry analysis and long-term cell proliferation analysis. For these experiments, we generated stable Flp-In 293 T-Rex cells with inducible expression of both wild-type and mutant ZBTB2 and PSRC1. No construct was available for RANBP2, due to its large size (~ 12kb). Mass spectrometry analysis of ZBTB2 and PSRC1, both wild-type and mutant forms, was performed to systematically map all protein interactions. The protein interactomes consisted of several known cancer-associated proteins (data not shown) and proteins with molecular functions relevant to tumorigenesis (see Figure 6). Moreover, the hot spot mutations in ZBTB2 and PSRC1 were shown to alter the protein interactomes (see Supplementary Figures 2C and 2D in the original publication).

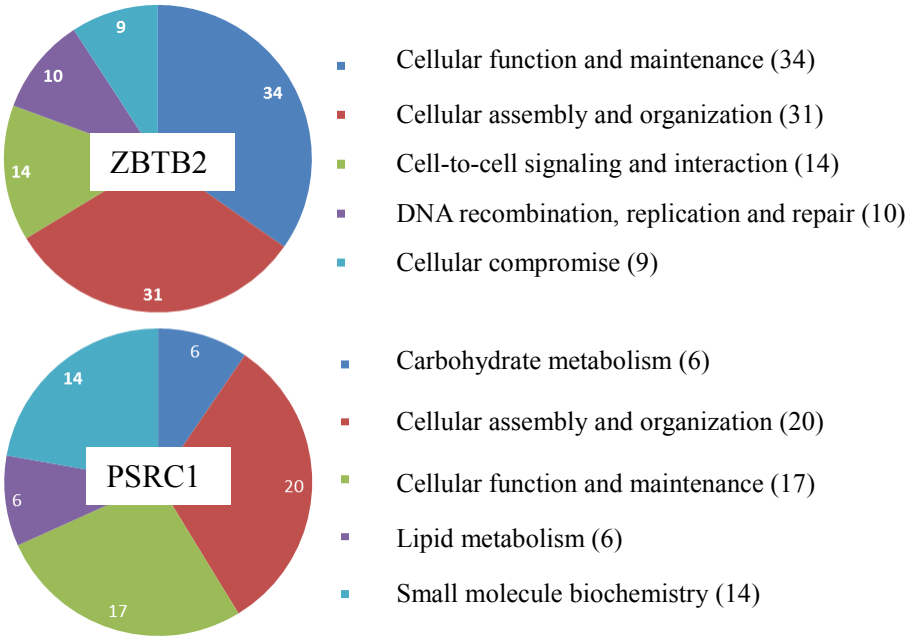

Figure 6. Interactomes of ZBTB2 and PSRC1. The pie diagrams show the distribution of ZBTB2 and PSRC1 interaction proteins according to their top five molecular and cellular functions. Numbers indicate the number of interaction proteins in each group. Data on protein functions were obtained with Ingenuity Pathways Analyses software.

Next, we wanted to study whether the mutant proteins caused an increase in cell proliferation. Results from the long-term cell proliferation assay showed a clear increase in proliferation in cells with the hot spot mutation (Arg261Trp) in ZBTB2 (Figure 7). No difference in proliferation was observed between wild-type and

mutant forms of PSRC1.

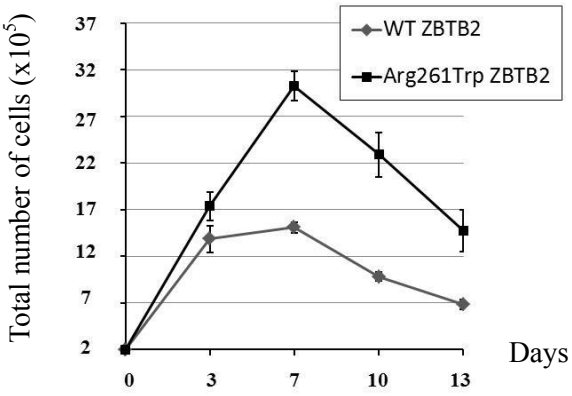

Figure 7. Mutant ZBTB2 showed increased cell proliferation. An increase in cell proliferation was observed for the CRC-associated mutant form of ZBTB2 (Arg261Trp) in HEK293 cells. Error bars depict the standard error of the mean.

3 Mutational profiles of fifteen candidate cancer genes in familial colorectal cancer

A systematic sequencing study conducted by Wood *et al*. (2007) identified 140 cancer candidate genes (CAN genes) somatically mutated in CRC. In our effort, the entire coding regions of 15 top-ranked CAN genes (see Table 6) were screened for somatic mutations and germline variants in 45 familial CRC cases. Six of the genes were

Table 6. Mutational profiles of fifteen CAN genes in familial CRC

Gene	Somatic mutations		Germline variants**	Population matched controls
	Mutation frequency*	Mutation types		
PIK3CA	8/45, 18%	Missense		
FBXW7	8/45, 18%	Missense (7) Nonsense (1)		
CSMD3	2/45, 4.5%	Missense	c.4045T>G, p.F1349V	0/865
TNN	0/45			
NAV3	0/45			
EPHA3	0/45			
MAP2K7	0/45			
EPHB6	1/45, 2%	Missense	c.961G>C, p.A321P	0/843
ADAMTSL3	0/45			
GUCY1A2	0/45			
SMAD2	1/45, 2%	Missense		
OR51E1	0/45			
LAMA1	0/45			
c10orf137	0/45		c.827T>C, p.I291T	0/876
TCF7L2	2/45, 4.5%	Missense, Splice site		

* Silent mutations were not included in the counts

**Only germline variants that were absent in controls are listed here

somatically mutated in the familial CRC cases with a total of 22 non-synonymous mutations. The most frequently somatically mutated genes were *phosphatidylinositol-4,5-bisphosphate 3-kinase catalytic subunit alpha* (*PIK3CA*, 18%) and *F-box and WD repeat domain containing 7 E3 ubiquitin protein ligase* (*FBXW7*, 18%) with most mutations locating at previously reported mutation hot spots. In the germline, three novel missense variants were identified in *CUB and Sushi multiple domains 3* (*CSMD3*), *chromosome 10 open reading frame 137* (*C10orf137*) and *EPH receptor B3* (*EPHB3*). None of them were present in the 890 population-matched controls (Table 6). The respective tumors did not show LOH. Only the variant in *C10orf137* was shown to segregate in the family. Currently, the functions of *CSMD3* and *C10orf137* have been poorly defined, and further studies are required to clarify their potential role in cancer. *EPHB3* encodes for an ephrin receptor, which belongs to the Eph receptor tyrosine kinase family, and has been suggested to act as a tumor suppressor in CRC tumorigenesis (Battle et al., 2005).

4 Eleven novel candidate susceptibility genes for familial colorectal cancer

Exome sequencing was conducted on germline DNA from 96 independent familial CRC cases. As part of previous efforts (Aaltonen et al., 1998; Salovaara et al., 2000), all cases had been shown to be mutation negative for any known high-penetrance CRC predisposing gene by standard clinical and molecular approaches. To further exclude such cases, the exome data was first analyzed for mutations in the following known high-penetrance CRC genes: *MLH1, MSH2, MSH6, PMS2, APC, MUTYH, SMAD4, BMPR1A, STK11* (also known as *LKB1*), *PTEN, AXIN2, POLE* and *POLD1.* Exome sequencing revealed no clear predisposing mutation in these genes. Next, we searched for novel candidate predisposing genes with rare putative LoF variants (nonsense, frameshift and splice-site). A detailed presentation of the steps taken in this study and the number of variants at each step is presented in Figure 8.

In total, we identified 11 novel candidate predisposing genes with putative LoFs (Table 7). Nine genes showed LoFs in 2/96 familial cases and two genes showed LoFs in 3/96 familial cases. All variants had a MAF of ≤ 0.001 in Finnish population matched controls (Table 7). We also examined variant frequencies in Exome variant server (http://evs.gs.washington.edu/EVS/) and 3/14 variants were found, however at very low frequency (MAF < 0.0003). This data set was not available at the time of data analysis and was therefore not included in the initial control filtering step. Interestingly, the splice-site variant in *PRADC1* was also identified in a nonsyndromic Finnish CRC patient diagnosed at the age of 36 years (unpublished data). The patient had no first-degree relatives with CRC. The candidate CRC predisposing loci were also searched for missense variants. Five missense variants, in five genes, were identified and these were all rare in population matched controls

(MAF < 0.001) (Table S3 in the original publication). None of them were predicted to have a damaging effect on protein function by either of the prediction programs used.

Table 7. Eleven novel candidate susceptibility genes for familial CRC

		Discovery phase samples		Validation phase samples			
Gene	**Variation (Amino acid)***	Familial cases	Finnish population matched controls**	Finnish population matched cases	Finnish population matched controls	**Loss of wt allele**	**Segregation**
UACA	p.Q1116X	2/96	1/522	2/862	1/550	3/4	
UACA	p.QR1292X	1/96	0/494	1/823	0/550	0/2	
SFXN4	fs	3/96	1/502			0/3	yes
TWSG1	p.Q41X	2/96	0/494	0/886	0/545	2/2	partial
PSPH	fs	2/96	1/502			1/2	
NUDT7	p.Y37X	2/96	0/494			0/2	
ZNF490	p.R350X	2/96	0/491	1/877	0/551	1/3	no
PRSS37	sp	1/96	0/491			0/1	
PRSS37	p.W138X	1/96	0/489			0/1	
CCDC18	sp	1/96	0/492			0/1	
CCDC18	p.S1109X	1/96	0/475			0/1	
PRADC1	sp	2/96	0/482			0/2	yes
MRPL3	sp	2/96	0/487			0/2	
AKR1C4	fs	2/96	0/491			0/1	yes

**fs* = frameshift insertion and deletion variant, sp= splice site variant

** Control counts include both exome data and Sanger sequenced controls

Loss of heterozygosity (LOH) was examined in the tumor tissue of cases carrying candidate predisposing variants. Altogether, seven LOH events were observed (Table 7) and the wild-type allele was lost in all seven occasions (P= 0.0078). Variants in genes showing LOH in the tumor tissue were genotyped in an independent population matched set of 954 CRC cases and 586 controls. Genotyping results are presented in Table 6, in the column termed validation phase samples. Genotyping was not successful for c.389_390insA in *PSPH*. Segregation analysis was performed whenever possible and results are presented in Table 7 and pedigrees are shown in the original publication (Figure 3 and Figure S1).

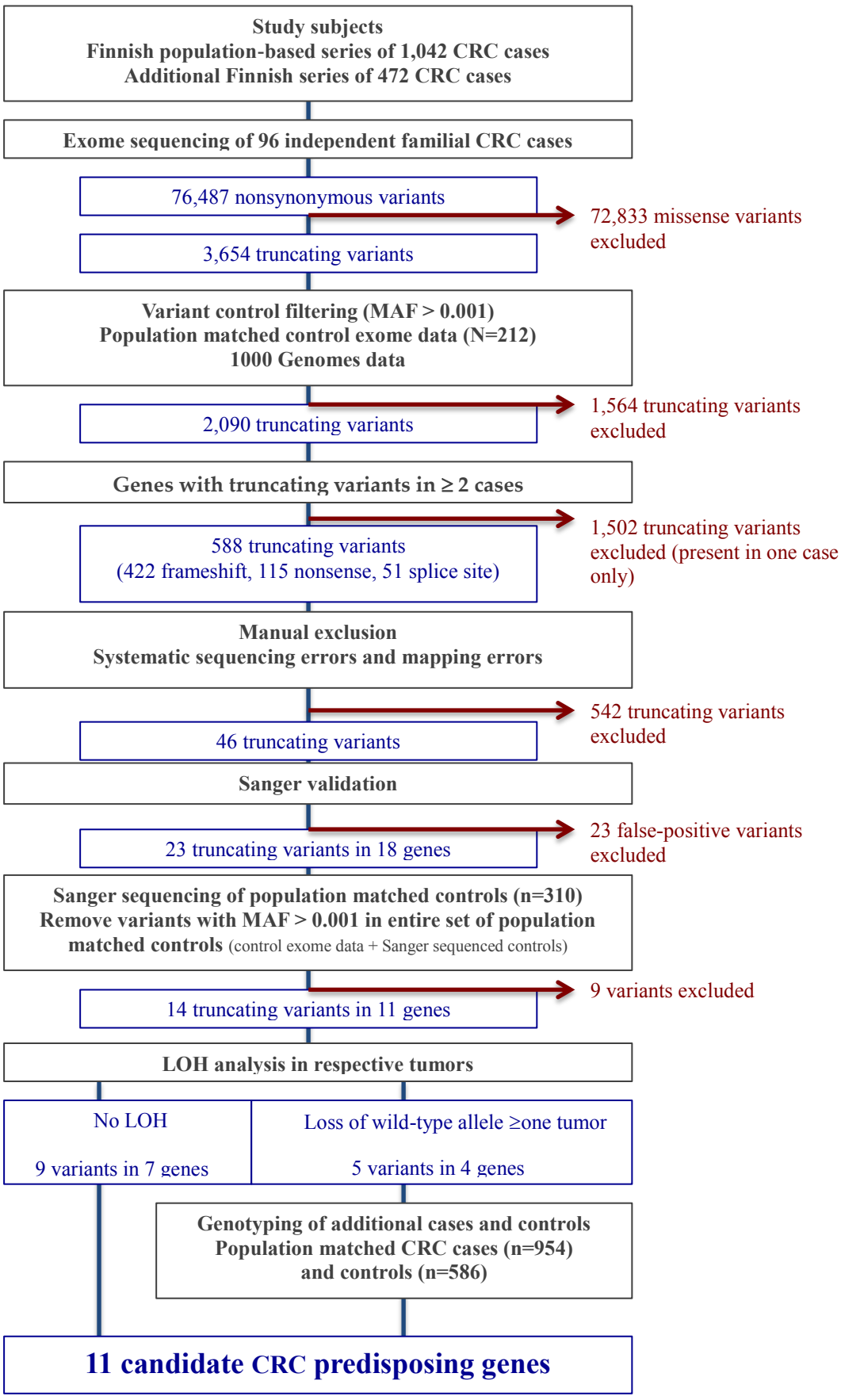

Figure 8. Summary of the overall study design used in study IV. MAF = minor allele frequency

DISCUSSION

1 *TTK* mutations in microsatellite-unstable colorectal cancer

MSI tumorigenesis is driven by a defective mismatch repair system and results in accumulation of frameshift mutations at microsatellite repeats (Aaltonen et al., 1998; Salovaara et al., 2000). Many genes have been suggested as MSI targets, usually based on high coding repeat frequency, which is thought to be one of the most important criteria for identifying true MSI driver genes (Boland et al., 1998). Background mutation frequencies are, however, extremely high in this tumor type and should always be taken into account when analyzing MSI target genes. We have previously generated reference datasets on frameshift mutations occurring in non-coding repeats to gain better insight into background mutation rates (Sammalkorpi et al., 2007; Alhopuro et al., 2012). Coding frameshift mutations at repeats typically lead to down-regulation of the target genes due to nonsense-mediated mRNA decay (Conti & Izaurralde, 2005). However, frameshift mutations located at the 3' end of the coding regions are often missed by the NMD machinery (Nagy & Maquat, 1998; MacArthur et al., 2012). Examples of MSI target genes with such frameshift mutations are *BAX* and *TCF7L2* (El-Bchiri et al., 2005). In this study, we sought to identify novel MSI target genes with putative oncogenic functions. We systematically characterized genes with NMD-decay escaping frameshift mutations that were overexpressed in MSI CRCs.

We identified frequent (59%, 105/179) decay-escaping mutations in the mitotic checkpoint kinase *TTK* (also known as *MPS1*). Repeat mutations in the last exon of *TTK* have been previously reported in 27-52% of MSI tumors. However, these studies have included a fairly limited set of tumors (Mori et al., 2002; Ahn et al., 2009; Kim et al., 2013). In our data, we also found enrichment of exon 5 mutations in exon 22 wild-type tumors. Based on our mutational data, *TTK* mutations appear to be selected for during tumor evolution, since the mutation frequency is significantly higher under mismatch repair deficiency than observed in identical control repeats. In addition, we were able to show the presence of the 34 residues elongated mutant protein in heterozygous cell lines, confirming that the mutant mRNA actually does escape NMD. Moreover, the mutant form of the protein was shown to locate normally to prometaphase kinetochores, which was somewhat expected since kinetochore localization has been shown to be mediated by residues in the N-terminus of TTK (Liu et al., 2003).

The TTK kinase plays important roles in mitotic regulation and spindle-assembly checkpoint (SAC) signaling. The SAC checkpoint ensures that chromosome segregation occurs correctly. TTK locates to kinetochores where it together with other checkpoint components, such as BUB1B and MAD2, delays anaphase onset

until all sister chromatids are properly attached to microtubules of the mitotic spindle. Major SAC defects have shown to lead to mitotic catastrophe and ultimately to apoptosis, whereas minor defects may result in aneuploidy (Kops et al., 2005). Several SAC components have shown to be mutated in CIN CRCs, such as BUB1 and BUB1B. TTK is also involved in other processes such as centrosome duplication and cytokinesis (Fisk et al., 2003).

Since the main function of TTK is to regulate SAC, we wanted to investigate if the identified TTK mutations in MSI CRC alter the checkpoint. Mutant MSI CRC cell lines were challenged with paclitaxel to test their checkpoint arrest. Paclitaxel is a microtubule-stabilizing drug that induces SAC by affecting microtubule dynamics and tension (Swanton et al., 2007). Our results showed no evidence for SAC impairment in TTK mutant cells. Previous studies have shown the C-terminus of TTK to be important for SAC arrest, however, the c.2560delA mutation that elongates the C-terminus did not cause SAC override in the cells (Sun et al., 2010).

TTK is also known to have other cancer-related functions, in addition to SAC regulation, that could potentially be affected by the identified mutations. For instance, substrate recruitment has been shown to be mediated by the C-terminal region of TTK and might be altered by the protein-elongating mutations (Sun et al., 2010). Previous studies have shown TTK to phosphorylate proteins such as BLM and Smad2 and Smad3 (Leng et al., 2006; Zhu et al., 2007). Future studies will also be required to determine if the mutations affect the kinase activity of the TTK protein. In addition, the oncogenic BRAF (V600E) mutant has been shown to phosphorylate TTK, thereby preventing its degradation in melanoma (Liu et al., 2013). Mutated BRAF is common in sporadic MSI CRC, but we found no correlation between BRAF (V600E) and TTK mutations. Although no functional evidence of oncogenic mechanisms for mutant TTK was observed, the high mutation frequency combined with the mutation pattern argues for biological significance. Future studies will be needed to uncover the effects of the identified TTK mutations in CRC-related functions outside SAC.

2 Oncogenic mutations in microsatellite-unstable colorectal cancer

Today, only few oncogenes have been implicated in MSI tumorigenesis. Well-known examples are *BRAF, CTNNB1, PIK3CA* and *KRAS* (Rajagopalan et al., 2002; Shitoh et al., 2001). Oncogenes usually show highly characteristic and non-random mutation patterns, with missense mutations at recurrent positions. *BRAF* is mutated in approximately one-third of MSI tumors and almost always at the codon V600E (Rajagopalan et al., 2002). Vogelstein *et al*. (2013) recently stated that one of the best way to identify driver genes mutated in cancer is through analyzing their mutation pattern rather than their mutation frequency. In general, MSI tumors are difficult to study since they harbor an extremely high number of passenger mutations. We

hypothesized that by searching for very specific mutation patterns we can identify true driver oncogenes that increase the selective growth advantage of MSI tumor cells.

To identify novel MSI CRC oncogenes, we sequenced the exomes of 25 MSI tumors and respective healthy tissue. We identified 15 novel candidate oncogenes recurrently mutated at the same amino acid positions. *ZBTB2, RANBP2* and *PSRC1* were our top hits, since they showed additional hot spot mutations in the validation set. The overall frequencies of hot spot mutations were 3.8% (4/106 MSI CRCs), 3.7% (4/109) and 4.6% (5/109) for *ZBTB2*, *RANBP2* and *PSRC1*, respectively. The mutation frequencies observed are not that modest, given our strict criterion for hot spots; missense mutations had to locate at the same or adjacent codons. For comparison, well-known oncogenes such as *KRAS* and *CTNNB1* had hot spot mutation frequencies of 8% in the exome data.

Hot spot mutations in our top three genes have not been reported in previous systematic sequencing efforts on CRCs. However, none of the studies have had a particular focus on MSI CRCs (Seshagiri et al., 2012; Cancer Genome Atlas Network, 2012). Of note, two other missense mutations, which were also found in our data, were reported in *RANBP2* and *ZBTB2*. These mutations located close to our hot spot site. In our study, the hot spot mutations in the three top genes were exclusively found in MSI CRCs; however, to confirm this finding a larger sample set needs to be screened. Interestingly, also in the TCGA data a mutation at our hot spot site in *ZBTB2* was found in a gastric tumor (MSI status not available, http://www.cbioportal.org/public-portal/). The hot spot mutations should be screened for in a large set of cancers with MSI, such as endometrial and gastric cancers.

2.1 Function of ZBTB2, RANBP2 and PSRC1 in health and disease

The transcription factor ZBTB2 has recently been reported as a central regulator of the TP53 pathway, through repressing transcription of *ARF, TP53* and *P21* and activating transcription of *HDM2*. HDM2 is suggested to induce rapid degradation of TP53, which further blocks the protective effect of TP53 in cellular response to DNA damage. In our study, the CRC-associated mutant form of ZBTB2 showed an increase in long-term cell proliferation. The previous finding that knock-down of ZBTB2 decreased cell proliferation (Jeon et al., 2009) further supports an oncogenic role for ZBTB2. Based on protein function, ZBTB2 serves as an excellent target for activating oncogenic mutations.

RANBP2 (also known as NUP358) is a large nucleoporin, a component of the nuclear pore complex, consisting of several domains that each interact directly and

selectively with different proteins, the RAS-related GTPase RAN being one of them. The interacting proteins all show distinct cellular or molecular functions, highlighting the pleiotropic role of RANBP2 (Vetter et al., 1999). Previous studies have reported RANBP2 to act as a tumor suppressor by altering chromosomal instability by regulating topoisomerase IIα by sumoylation (Navarro & Bachant, 2008). *RANBP2* is over-expressed in several types of cancers, including CIN CRC cell lines, multiple myeloma and mouse prostate cell lines (Dunican et al., 2002; Felix et al., 2009; Renner et al., 2007). Interestingly, another study has suggested RANBP2 to be involved in CRC tumorigenesis by regulating WNT signaling. Over-expression of the protein increased nuclear import of TCF-4 and β-catenin, which enhanced transcriptional activity and resulted in increased growth of CRC cell lines (Shitashige et al., 2008). It is hence possible that the mutated RANBP2, identified here, acts as an oncogene by enhancing the transcriptional activity of β-catenin and TCF4, thus inducing activation of the WNT signaling pathway.

PSRC1 is a microtubule-binding and bundling protein, the expression of which has shown to be elevated in several cancers, supporting an oncogenic role in tumorigenesis (Hsieh et al., 2002; Sun et al., 2008). PSRC1 has been linked to inhibition of TP53-mediated apoptosis, via inhibition of ASPP2, a regulator of apoptosis and cell growth. The ASPP2-interaction domain of PSRC1 has been mapped to amino acids 118-141, covering the hot spot site identified in our study (Sun et al., 2008). However, mass spectrometry data did not reveal an altered binding to ASPP2 in the mutant cells. Interestingly, PSRC1 has also been shown to promote cell growth by enhancing β-catenin mediated transcriptional activation of the WNT signaling pathway (Hsieh et al., 2007). However, we observed no evidence that the hot spot mutations would cause increased cell proliferation in long-term cell cultures.

Much experimental work will be needed to elucidate the detailed functional consequences of the identified mutations. Nevertheless, the mutation pattern, missense mutations at recurrent locations, combined with the functional roles of these three genes provides strong evidence for oncogenic effects in tumorigenesis. Driver oncogenes with activating mutations, located at the same amino acid positions, are attractive targets for small-molecule inhibitor drugs. For example melanoma patients harboring *BRAF* V600E mutations are already successfully treated with specific kinase inhibitors, which result in dramatic tumor remission (Chapman et al., 2011). However, as with all cancer therapies, there are some immediate limitations that need to be addressed. One challenge is that pathway functions are different in different tissues and affected by the accompanied genetic alterations. The same drugs that were used to treat melanoma patients have no therapeutic effect in CRC patients with *BRAF* V600E mutations (Mao et al., 2013). Also, only a few tumors harbor more than one mutated oncogene, which is a challenge for combination strategies in targeted therapies. Nevertheless, there is an

urgent need to better understand cancer pathways and driver genes behind cancer development. Ultimately, this knowledge will guide the development of more effective approaches in cancer treatment.

In the near future, it is estimated that tens of thousands of cancer genomes will be sequenced. Large-scale studies will most likely be done by utilizing whole-genome sequencing, combined with data on the transcriptome and epigenomes from the same cases. Ultimately, these studies will have a great impact on our understanding of cancer biology and unravel attractive new strategies for therapies and prevention.

3 Susceptibility genes for common familial colorectal cancer

It is well known that family history is one of the strongest risk factors for the development of CRC. Many CRCs develop in genetically susceptible individuals, most of whom are not carriers of high-penetrance gene mutations, such as *APC* or mismatch repair mutations (Lichtenstein et al., 2000; Aaltonen et al., 2007; Lubbe et al., 2009). Previous studies have reported familial CRC to be fairly common, with approximately one tenth of CRC patients having a family history of the disease, if excluding high-penetrance mutation carriers (Aaltonen et al., 2007). First-degree relatives of CRC patients show a two-fold increase in risk compared to the general population (Johns & Houlston, 2001). Despite extensive efforts, the molecular background of familial risk remains unexplained. An important part might be explained by unknown rare variants of moderate-penetrance.

3.1 The role of fifteen candidate cancer genes

Sjöblom et al. in 2006 and Wood et al. in 2007 performed the first systematic sequencing efforts on nearly all protein-coding genes in breast and colorectal cancer. The study by Wood *et al*. (2007) included a discovery set of 11 breast and 11 colorectal cancers. By utilizing a statistical tool that took into account background mutation frequencies, 140 somatically mutated candidate cancer (CAN) genes were reported, for both cancer types. The CAN genes were ranked based on the likelihood of observed gene mutation prevalence exceeding the expected background mutation prevalence (Wood et al., 2007). We know from previous studies that somatically mutated genes are often involved in hereditary predisposition (Futreal et al., 2004) and the novel CAN genes thus serve as obvious candidates for CRC predisposition. We analyzed the mutational profile of 15 top-ranked CAN genes (Wood et al., 2007) in 45 familial CRC cases for both somatic mutations and germline variants.

Our data on somatic mutations in CAN genes were in good agreement with previous studies (Wood et al., 2007; Sjöblom et al., 2006). As expected, the top-ranked CAN genes *PIK3CA* and *FBXW7* were the most frequently mutated genes in our sample set. *PIK3CA* encodes for the protein p110α, which is the catalytic subunit

of the class I PI3-kinase. The gene is a known oncogene with mutations that lead to constitutive activation of the PI3K pathway in several common cancers. It has been reported that the mutations are clustered and locate mostly in the helical and kinase domains of the protein, which was further supported by our findings (Samuels & Velculescu, 2004; Ikenoue et al., 2005). FBXW7 is one of the four subunits of the ubiquitin ligase complex, which functions in phosphorylation-dependent ubiquitination of proteins, such as MYC and cyclin E (Koepp et al., 2001; Yada et al., 2004). FBXW7 functions as a tumor suppressor, and recent studies have also suggested a dominant-negative role of the mutations in cancer development (Welcker & Clurman, 2008).

In this effort, we did not identify a clear CRC predisposing gene. Nevertheless, three novel non-synonymous germline variants were found in *CSMD3*, *EPHB3* and *C10orf137,* and none of them were present in the 890 population-matched healthy controls. Additional studies with larger sample sets will be required to clarify the role of these variants in CRC predisposition.

3.2 Identification of susceptibility genes by exome sequencing

Improvements in sequencing technologies have provided novel tools to study cancer predisposition in an unbiased manner. In this effort, we sought to identify novel susceptibility genes for common familial CRC. CRC families with few affected individuals are an attractive patient group to search for novel susceptibility genes, but tools for such work have been poor. Here**,** we performed exome sequencing on 96 independent familial CRC cases derived from a constitutive collection of unselected patients (Aaltonen et al., 1998; Salovaara et al., 2000). All cases were from Finland, with a population known for its relatively uniform genetic background. In principle, this facilitated gene identification since individuals from isolated populations are more likely to share ancestral predisposing mutations, originating from a few common founders (Peltonen et al., 2000; Jakkula et al., 2008). In this study, we identified 11 genes with rare truncating variants in two or three familial CRC cases. They were all absent or rare (MAF ≤ 0.001) in the general Finnish population. None of the novel candidate susceptibility genes had been previously implicated in cancer predisposition. Proposed gene functions and suggested pathways in which the encoded proteins are involved are presented in Table 8.

Our effort focused solely on a subset of variants: rare variants with MAF ≤ 0.001 in the general population. The study was, thus, conducted under the "rare variant hypothesis" that proposes that a significant portion of the missing heritability is due

Table 8. Eleven candidate susceptibility genes for common familial CRC*

Gene	Gene Description	Protein function	Biological process	Canonical pathway
UACA	*uveal autoantigen with coiled-coil domains and ankyrin repeats*	protein binding	apoptosis	
SFXN4	*sideroflexin 4*	transporter	transmembrane transport	
TWSG1	*twisted gastrulation BMP signaling modulator 1*	protein binding	cell differentation, apoptosis	TGF-β signaling
PSPH	*phosphoserine phosphatase*	phosphatase	amino acid biosynthesis	serine biosynthesis
NUDT7	*nudix (nucleoside diphosphate linked moiety X)-type motif 7*	acetyl-CoA hydrolase	cell differentation, acetyl-CoA catabolic process	
ZNF490	*zinc finger protein 490*	DNA binding	regulation of transcription	
PRSS37	*protease, serine, 37*	protease	proteolysis	
CCDC18	*coiled-coil domain containing 18*	not known	not known	
PRADC1	*protease-associated domain containing 1*	not known	not known	
MRPL3	*mitochondrial ribosomal protein L3*	RNA binding, structural constituent of ribosome	translation	
AKR1C4	*aldo-keto reductase family 1, member C4*	reductase	metabolic processes	androgen and estrogen biosynthesis

* Data on protein functions, biological processes and canonical pathways obtained with Ingenuity Pathways Analyses software (www.ingen.com). Only selected protein functions and processes are listed here.

to a series of rare variants of moderate-penetrance. However, expressivity may be altered by other loci or environmental factors (Fletcher & Houlston, 2010; Bodmer & Bonilla, 2008). Evolutionary theory strongly supports this hypothesis by arguing that disease promoting variants are selected against and must therefore be rare in the population. The hypothesis is further supported by recent empirical population genetic data, such as the data from the 1000 Genomes project, which reported rare variants to be enriched for deleterious mutations (Gibson, 2011; 1000 Genomes Project Consortium et al., 2012). Rare variants are usually population specific and for this reason it was important to filter the identified variants against a large number of

population matched healthy controls to further exclude neutral polymorphisms (1000 Genomes Project Consortium et al., 2012).

In this study, focus was placed on variants predicted to truncate the protein product, so called LoF variants, which are attractive candidates for disease predisposition. Previous studies have shown true LoF variants, to be mostly of low frequencies in the population. They tend to be mildly or severely deleterious, and have therefore been stopped by natural selection from increasing in frequency (MacArthur et al., 2012; 1000 Genomes Project Consortium et al., 2012). However, recent systematic sequencing studies have revealed a surprisingly high number of LoFs in healthy individuals, most of which have little or no effect on health. This further highlights the challenge to identify true predisposing LoFs from the background of non-pathogenic polymorphisms. In addition, analysis on gene expression has revealed an unanticipated high amount of LoFs to be neutral, with no effect on gene function (MacArthur et al., 2012). Further studies are needed to clarify if the LoFs identified here are actually genuine LoFs that cause reduction in gene expression.

Also, the question remains whether the identified candidate genes act as classical tumor suppressors. Seven LOH events were observed, involving four of the candidate genes. Interestingly, in all seven occasions the wild-type allele was lost (P=0.0078), which provides us with additional evidence that true predisposing genes are among the eleven genes. In at least a subset of the identified genes, complete inactivation seems to be preferentially selected for in tumor evaluation. However, it is plausible that some of the candidate genes show alternative mechanisms, such as haploinsufficiency or dominant-negative effects. In summary, we have identified an interesting set of candidate predisposing genes that may explain a subset of common familial CRC. However, additional genetic validation in larger sample sets, representing different populations, is required to gain robust evidence for pathogenicity. Also, assessing the functional effects of the identified LoFs in tumor development and progression would be a crucial next step.

Exome sequencing is an attractive approach to study susceptibility to cancer, in which coding germline variants, at the entire allelic spectrum, can comprehensively be investigated. However, studies on rare predisposing variants are at the moment largely underpowered. It has been suggested that as many as 10.000 exomes are needed to achieve sufficient statistical power to robustly detect associations of rare variants with complex traits (Kiezun et al., 2012). Also, associations need to be replicated in independent sample sets from different populations. The cost of sequencing is falling at a dramatic pace. Exome sequencing, as well as WGS, will in the near future be affordable to many research groups. Pooling sequencing data, through established consortia, will enable the generation of well-powered experiments that will lead to novel discoveries.

Noncoding genetics variants, located at regulatory regions, several forms of structural variations and polygenic inheritance, have been largely unexplored as a basis for CRC susceptibility. Future studies, utilizing WGS, will likely uncover novel variants of this class.

CONCLUSIONS AND FUTURE PROSPECTS

This study was conducted to provide novel insight into the molecular genetic background of CRC. Studies I and II focused on somatic mutations in MSI CRC and studies III and IV on germline variants underlying familial CRC predisposition.

I-II

MSI is characteristic for Lynch syndrome and observed in a subset of sporadic CRC cases. Until today, only few unbiased large-scale sequencing efforts on MSI CRC have been conducted. In addition to many target genes with inactivating mutations, only a handful of oncogenes have been implicated in this tumor type. The aim of our efforts was to identify novel oncogenes in MSI CRC. In study I, we systematically characterized NMD-escaping target genes, overexpressed in MSI tumors. We identified frequent frameshift mutations in the mitotic checkpoint kinase *TTK* that resulted in an elongated protein. When compared to background frequencies, *TTK* was found to have significantly higher mutation frequencies than expected without clonal selection. By *in vitro* functional assays, no evidence of oncogenic mechanism was observed; however, the high mutation frequency of *TTK* argues for selection in tumorigenesis. In study II, exome sequencing of 25 MSI CRC tumor-normal pairs revealed 15 novel candidate oncogenes with hot spot mutations. *ZBTB2*, *RANBP2* and *PSRC1* showed hot spot mutations also in the validation set. Corroborating previous data, our results from functional studies on these three genes suggest a role in cancer development and progression. Additional efforts are needed to fully understand the nature and functional significance of the identified somatic mutations in CRC tumorigenesis.

The findings of studies I and II further underline the notion that CRC genomes are heterogeneous, characterized by few frequently mutated genes, called mountains (e.g. *BRAF* and *PIK3CA*), and numerous less frequently mutated genes, called hills (e.g. *NRAS* and here-identified *PSRC1*, *RANBP2* and *ZBTB2*) (Wood et al., 2007; Sjöblom et al., 2006). Creating a comprehensive catalogue of all CRC genes mutated at intermediate frequencies, including those identified here, is crucial in order to recognize dysregulated pathways and optimal targets for therapeutic intervention. In addition, such a catalogue would lead us one step closer to personalized medicine, where the choice of combination therapy for each individual patient would be based on the cellular pathways dysregulated in their tumor.

Oncogenes with activating hotspot mutations, such as *ZBTB2*, serve as attractive targets for therapeutic interventions. Cancer genome sequencing efforts have already had an impact on the clinical care of cancer patients. For example, the identification of activating mutations in genes encoding protein kinases has led to the

development of small-molecular inhibitors targeting those kinases. Such approaches include EGFR kinase inhibitors to treat cancers with *EGFR* mutations, such as gefitinib for non-small-cell lung cancer (Sharma et al., 2007), and specific inhibitors of mutant BRAF to treat cancers with *BRAF* mutations, such as vemurafenib for metastatic melanoma (Chapman et al., 2011). In addition, identification of novel cancer genes, such as *TTK* with protein-elongating mutations, may have implications for diagnostics and for cancer immunotherapies by providing novel tumor-specific antigens. Circulating autoantibodies generated by the immune cells against tumor-specific antigens can serve as detection patterns of malignancy for earlier cancer diagnosis and predicting outcomes (Casino et al., 2006). Moreover, such antigens could be used in already existing platforms for immunotherapy of cancers, such as vaccines containing the immunogenic mutant peptides, antibodies developed against the tumor-specific antigens, or T cells with reactivity directed against the mutant proteins (Kirkwood et al., 2011).

In the near future, as more tumors are being sequenced, it will be fairly straightforward to identify all genes mutated at elevated rates in CRC. However, characterization of their functional role in tumorigenesis will be a crucial, yet challenging task. Furthermore, studies to characterize and interpret alterations in noncoding DNA, DNA methylation, mRNA expression and protein expression in CRC are only in their early stages.

III-IV

Hereditary factors are presumed to play a major role in CRC risk; however, still today the etiology of familial CRC is largely unknown. This is particularly the case for common CRC families with few affected individuals. In study III, fifteen top-ranked somatically mutated CRC genes, previously published by Wood *et al*. (2007), were screened for germline mutations in familial CRC cases. None of the genes were shown to clearly predispose to familial CRC; however, three novel missense variants were identified that were absent in population matched controls. In study IV, we exome sequenced 96 independent familial CRC cases, with typically only one affected first-degree relative. We identified eleven novel candidate predisposing genes with rare protein-truncating variants in familial CRC cases. Seven LOH events were observed in the respective tumors and in all occasions the wild-type allele was lost, which provides us with additional evidence that true culprits are among these genes.

The architecture of inherited genetic susceptibility to CRC is complex: characterized by multiple predisposition alleles with different levels of risk and prevalence in the population. Our results expand on the existing repertoire of genes that might predispose to CRC. Further investigations, including genetic validation in large sample sets representing different populations, are required to provide robust

evidence for disease causality. Additional work is also needed to characterize the detailed functional and clinical relevance of the identified candidate CRC predisposing genes.

In the next few years, exome sequencing-efforts with large sample sets will most likely identify many more variants associated with CRC. A future challenge to be faced might not lie in the identification of the association signals alone, but rather in characterizing the molecular mechanisms in which they influence cancer risk. Such insights will provide greater understanding of cancer biology in general and reveal potential targets for diagnostic and therapeutic strategies. Also, such information would be of immediate clinical relevance in regard to cancer risk assessment. The identification of individuals at increased risk allows for targeted cancer prevention strategies and can also influence cancer treatment options.

There is an urgent need for new technologies that would transform functional genomics in cancer research. Efficient large-scale functional studies will be required to ultimately characterize the genes and pathways important in cancer development. The functional consequences of putative LoF variants, including those identified in this study, can currently be investigated with, for instance, RNA interference and novel genome editing technologies, such as the promising RNA-guided CRISPR-Cas9 system (Wang et al., 2014). Results from such approaches could further guide the construction of animal model experiments that are crucial in understanding processes of human pathophysiology.

Exome sequencing has its limitations in studying the genetic architecture, since noncoding variants, for instance at regulatory regions, are missed, while being likely important for cancer predisposition. Moreover, currently little is understood of the polygenic basis and gene-environment interactions behind CRC predisposing. Still today, much of the inherited predisposition to CRC remains unaccounted for, thus highlighting the need for further efforts. Ultimately, improved knowledge of CRC predisposition will guide the development of more effective strategies for reducing CRC morbidity and mortality.

ACKNOWLEDGEMENTS

This thesis work was carried out at the Department of Medical Genetics, at Haartman Institute, and Research Program Unit in Biomedicum Helsinki, University of Helsinki during 2008-2014. Olli Jänne and Sampsa Hautaniemi, the former and present heads of Research Programs Unit, are warmly thanked for providing state-of-the-art research facilities and modern and pleasant working environments.

The Helsinki Biomedical Graduate Program is acknowledged for a first-class graduate education, financial support as well as the valuable educational experiences I got as a student council member. I have also been financially supported by: the Biomedicum Helsinki Foundation, the Finnish Cancer Society, Finska Läkarsällskapet and the Maud Kuistila Foundation. All are sincerely acknowledged.

My deepest gratitude is extended to my supervisor Lauri Aaltonen for his continuous support, flexibility and faith in me during these years. Lauri's curiosity, open-mindedness and thoughtful character have been a great inspiration to me. I admire his ability to build a fantastic team that consists of people who are passionate about science and keen to have fun. Despite his busy schedule he has always found time to discuss results, manuscripts and other matters related to PhD life. Tack Lauri! I am also extremely grateful to my other supervisor Auli Karhu for the every-day advice in the lab, for teaching me how to think critically as a scientist, for interesting discussions and silly laughs. I highly appreciate her support and for always believing in my work. My supervisors often conveyed two messages; "it's time to finish the thesis already", and "don't worry about when you'll finish, you'll get there eventually." They were wise enough to know when to say which.

Iiris Hovatta and Pipsa Saharinen are sincerely thanked for their in-depth review of the thesis manuscript and for providing me with valuable comments and thoughts that further strengthened this thesis. I owe many thanks to Tiia Pelkonen for excellent language editing of the thesis. I also extend my thanks to Iiris Hovatta (again) and Katri Koli for their efforts on the thesis committee and for their encouragements throughout the PhD.

I thank all collaborators without whom this work would not have been possible: surgeons Jukka-Pekka Mecklin, Heikki Järvinen and Laura Sinisalo-Renkonen for fantastic sample material, which has served as the key for successful research; Ari Ristimäki for pathological expertise; Markus Mäkinen and Torben Ørntoft for providing samples; Jussi and Minna Taipale for efforts on exome sequencing; Sampsa Hautaniemi for bioinformatics expertise and valuable advice at the initial stages of my scientific career; Marko Laakso and Ville Rantanen for skillful computational help and efforts on imaging; Marko Kallio, Leena Laine, and Jeroen Pouwels for their efforts on the TTK functional experiments; Mikko Turunen for much-needed help on the functional studies and for his continuous and selfless cooperation.

I want to express my warmest thanks to all members of the Aaltonen group, both new and old, for creating such a wonderful working environment. The post-docs and senior scientists in the lab are greatly acknowledged, in particular: Sari Tuupanen, without whom this thesis may have taken another 5 years, for her help, advice and encouragement throughout, for a laid-back family vacation in Fuerte (that I actually was not invited to) and becoming a trusted friend. To Heli Lehtonen, for her efforts on the functional studies and for great friendship – the kind that is hard to find. Thank you to all the other co-authors: Pia Vahteristo and Outi Kilpivaara, for all their help, kind support and positive spirit; Esa Pitkänen and Kimmo Palin, for their expertise in bioinformatics, easy-going attitude and good sense of humor.

I am indebted to my fellow PhD students in the close-knit CRC family – it has truly been a pleasure: I thank my dear CRC sister Iina Tuominen for her kind support during the first steps of my PhD journey and for warmly welcoming me into the *TTK* project; my close colleague Johanna Kondelin, for excellent help on basically anything that I could think to ask, for blending a cheerful personality with an incisive mind and for sharing my enthusiasm during these years – it's been refreshing; Tomas Tanskanen for endless interesting scientific discussions and for providing focus to my final project by pulling together all the loose threads; Riku Katainen, for his can-do attitude, Rikurator and help with figures – in a parallel universe you might be a brilliant artist; Tatiana Cajuso and Ulrika Hänninen for selfless help with the experiments and great attitude in the lab – what doesn't kill you makes you stronger! One never knows when a casual conversation leads one in a fruitful direction - so many thanks to the entire CRC family for all the interesting discussions.

My sincere gratitude also goes to Inga-Lill Svedberg, Mairi Kuris, Iina Vuoristo and Alison Ollikainen for excellent technical support and for taking care of everything in the lab. Warm thanks also to all other members of the Aaltonen lab: Sini Nieminen, Heikki Metsola, Marjo Rajalaakso and Sirpa Soisalo for scientific and practical help and for all their efforts in making the group a wonderful place to work; Mervi Aavikko, for enthusiastic and positive energy, great friendship and unforgettable ScienceSLAM moments; and all the above mentioned colleagues and the other members including Eevi Kaasinen, Heikki Ristolainen, Miika Mehine, Kati Kämpjärvi, Netta Mäkinen, Tuomas Heikkinen, Javier Nuñez Fontarnau, Rainer Lehtonen, Niko Välimäki, Iikki Donner, as well as former members of the lab, especially Manuel Ahlsten, Silva Saarinen, Elina Heliövaara, Jaana Tolvanen, Marianna Georgitsi and Anniina Raitila for making the group what it is and for so many countless occasions of fun, including sport tournaments, numerous parties, hiking in the wilderness and playing "Alias" in Swedish.

I wish to extend my thanks to all my Biomedicum friends: all student council members, including Sanna Vuoristo, Anna-Maria Lahesmaa-Korpinen, Irene Ylivinkka, Arvydas Dapkunas, Elina Hakonen and Katja Helenius for organizing unforgettable symposiums and events and for great parties; Viikki-friends Mervi Aavikko (again), Saija Ahonen, Sofia Ahola and Jenny Kivinen, for sharing thoughts and feelings about all sorts of aspects of PhD life and for fun nights at Puttes Pizza; Roxana Ola, Erkka Valo, Iulia Diaconu, Boris and Lotta Vassilev, Otto Hemminki, Ilkka Liikanen and Heli Lehtonen (again), for friendship, unforgettable moments and absolutely crazy times; including Hanko trips, beach volley games, climbing excursions and numerous karaoke-nights.

My sincere thanks go to my wonderful friends outside Biomedicum; Belina Björk, Anna Dahl, Linda Haglund, Annika Syrjänen, as well as other members of the girl gang. During these years they have kept me sane and in a bright and happy place. I'm truly grateful to have such wonderful friends! My very special thanks goes to Belina for always being there for me, for sharing my view of life and for being a fantastic person with a big heart and a great sense of humor.

Love and gratitude go to my dear family – my support team. My parents, Monica and Peter, are thanked for unwavering care, continuous encouragement and for always letting me choose my own path in life. Special thanks to my dad for lively discussions, wisdom and close friendship – none of this would have been possible without his support. Philip, Liisa and Ingrid (the new member) are thanked for lovely moments together. Also, my beloved Ilkka is thanked for his endless care and understanding – life with you is truly a joyful adventure!

Helsinki, May 2014

REFERENCES

1000 Genomes Project Consortium, Abecasis, G.R., Auton, A., Brooks, L.D., DePristo, M.A., Durbin, R.M., Handsaker, R.E., Kang, H.M., Marth, G.T. & McVean, G.A. (2012). An integrated map of genetic variation from 1,092 human genomes. *Nature*. 491 (7422). pp. 56–65.

Aaltonen, L., Johns, L., Järvinen, H., Mecklin, J.-P. & Houlston, R. (2007). Explaining the familial colorectal cancer risk associated with mismatch repair (MMR)-deficient and MMR-stable tumors. *Clinical cancer research.*13 (1). pp. 356–361.

Aaltonen, L.A., Peltomäki, P., Leach, F.S., Sistonen, P., Pylkkänen, L., Mecklin, J.P., Järvinen, H., Powell, S.M., Jen, J. & Hamilton, S.R. (1993). Clues to the pathogenesis of familial colorectal cancer. *Science*. 260 (5109). pp. 812–816.

Aaltonen, L.A., Salovaara, R., Kristo, P., Canzian, F., Hemminki, A., Peltomäki, P., Chadwick, R.B., Kääriäinen, H., Eskelinen, M., Järvinen, H., Mecklin, J.P. & la Chapelle, de, A. (1998). Incidence of hereditary nonpolyposis colorectal cancer and the feasibility of molecular screening for the disease. *The New England journal of medicine*. 338 (21). pp. 1481–1487.

Aarnio, M., Mecklin, J.P., Aaltonen, L.A., Nyström-Lahti, M. & Järvinen, H.J. (1995). Life-time risk of different cancers in hereditary non-polyposis colorectal cancer (HNPCC) syndrome. *International journal of cancer*. 64 (6). pp. 430–433.

Aarnio, M., Sankila, R., Pukkala, E., Salovaara, R., Aaltonen, L.A., la Chapelle, de, A., Peltomäki, P., Mecklin, J.P. & Järvinen, H.J. (1999). Cancer risk in mutation carriers of DNA-mismatch-repair genes. *International journal of cancer*. 81 (2). pp. 214–218.

Ahn, C.H., Kim, Y.R., Kim, S.S., Yoo, N.J. & Lee, S.H. (2009). Mutational analysis of TTK gene in gastric and colorectal cancers with microsatellite instability. *Cancer research and treatment*. 41 (4). pp. 224–228.

Al-Tassan, N., Chmiel, N.H., Maynard, J., Fleming, N., Livingston, A.L., Williams, G.T., Hodges, A.K., Davies, D.R., David, S.S., Sampson, J.R. & Cheadle, J.P. (2002). Inherited variants of MYH associated with somatic G:C-->T:A mutations in colorectal tumors. *Nature genetics*. 30 (2). pp. 227–232.

Alberici, P. & Fodde, R. (2006). The role of the APC tumor suppressor in chromosomal instability. *Genome dynamics*. 1. pp. 149–170.

Albuquerque, C., Breukel, C., van der Luijt, R., Fidalgo, P., Lage, P., Slors, F.J.M., Leitão, C.N., Fodde, R. & Smits, R. (2002). The 'just-right' signaling model: APC somatic mutations are selected based on a specific level of activation of the beta-catenin signaling cascade. *Human molecular genetics*. 11 (13). pp. 1549–1560.

Alhopuro, P., Sammalkorpi, H., Niittymäki, I., Biström, M., Raitila, A., Saharinen, J., Nousiainen, K., Lehtonen, H.J., Heliövaara, E., Puhakka, J., Tuupanen, S., Sousa, S., Seruca, R., Ferreira, A.M., Hofstra, R.M.W., Mecklin, J.-P., Järvinen, H., Ristimäki, A., Ørntoft, T.F., Hautaniemi, S., Arango, D., Karhu, A. & Aaltonen, L.A. (2012). Candidate driver genes in microsatellite-unstable colorectal cancer. *International journal of cancer*. 130 (7). pp. 1558–1566.

Amos C.I., Bali D., Thiel T.J., Anderson J.P., Gourley I., Frazier M.L., Lynch P.M., Luchtefeld M.A., Young A., McGarrity T.J., Seldin M.F. (1997). Fine mapping of a genetic locus for Peutz-Jeghers syndrome on chromosome 19p. *Cancer Research.* 1;57 (17) pp. 3653-6.

Andersen, C.L., Christensen, L.L., Thorsen, K., Schepeler, T., Sørensen, F.B., Verspaget, H.W.,

Simon, R., Kruhøffer, M., Aaltonen, L.A., Laurberg, S. & Ørntoft, T.F. (2009). Dysregulation of the transcription factors SOX4, CBFB and SMARCC1 correlates with outcome of colorectal cancer. *British journal of cancer*. 100 (3). pp. 511–523.

Araten, D.J., Golde, D.W., Zhang, R.H., Thaler, H.T., Gargiulo, L., Notaro, R. & Luzzatto, L. (2005). A quantitative measurement of the human somatic mutation rate. *Cancer research*. 65 (18). pp. 8111–8117.

Baker, S.J., Preisinger, A.C., Jessup, J.M., Paraskeva, C., Markowitz, S., Willson, J.K., Hamilton, S. & Vogelstein, B. (1990). p53 gene mutations occur in combination with 17p allelic deletions as late events in colorectal tumorigenesis. *Cancer research*. 50 (23). pp. 7717–7722.

Bamshad, M.J., Ng, S.B., Bigham, A.W., Tabor, H.K., Emond, M.J., Nickerson, D.A. & Shendure, J. (2011). Exome sequencing as a tool for Mendelian disease gene discovery. *Nature reviews. Genetics*. 12 (11). pp. 745–755.

Bansal, V., Libiger, O., Torkamani, A. & Schork, N.J. (2010). Statistical analysis strategies for association studies involving rare variants. *Nature reviews. Genetics*. 11 (11). pp. 773–785.

Barber, T.D., McManus, K., Yuen, K.W.Y., Reis, M., Parmigiani, G., Shen, D., Barrett, I., Nouhi, Y., Spencer, F., Markowitz, S., Velculescu, V.E., Kinzler, K.W., Vogelstein, B., Lengauer, C. & Hieter, P. (2008). Chromatid cohesion defects may underlie chromosome instability in human colorectal cancers. *Proceedings of the National Academy of Sciences of the United States of America*. 105 (9). pp. 3443–3448.

Batlle E., Henderson J.T., Beghtel H., van den Born M.M., Sancho E., Huls G., Meeldijk J., Robertson J., van de Wetering M., Pawson T., Clevers H. (2002). Beta-catenin and TCF mediate cell positioning in the intestinal epithelium by controlling the expression of EphB/ephrinB. *Cell*. 18 (2) pp. 251-63

Berger, A.H., Knudson, A.G. & Pandolfi, P.P. (2011). A continuum model for tumour suppression. *Nature*. 476 (7359). pp. 163–169.

Blanchette, M., Kent, W.J., Riemer, C., Elnitski, L., Smit, A.F.A., Roskin, K.M., Baertsch, R., Rosenbloom, K., Clawson, H., Green, E.D., Haussler, D. & Miller, W. (2004). Aligning multiple genomic sequences with the threaded blockset aligner. *Genome research*. 14 (4). pp. 708–715.

Bodmer, W. & Bonilla, C. (2008). Common and rare variants in multifactorial susceptibility to common diseases. *Nature genetics*. 40 (6). pp. 695–701.

Bodmer, W. & Tomlinson, I. (2010). Rare genetic variants and the risk of cancer. *Current opinion in genetics & development*. 20 (3). pp. 262–267.

Bodmer, W.F., Bailey, C.J., Bodmer, J., Bussey, H.J., Ellis, A., Gorman, P., Lucibello, F.C., Murday, V.A., Rider, S.H. & Scambler, P. (1987). Localization of the gene for familial adenomatous polyposis on chromosome 5. *Nature*. 328 (6131). pp. 614–616.

Boland, C.R., Thibodeau, S.N., Hamilton, S.R., Sidransky, D., Eshleman, J.R., Burt, R.W., Meltzer, S.J., Rodriguez-Bigas, M.A., Fodde, R., Ranzani, G.N. & Srivastava, S. (1998). *A National Cancer Institute Workshop on Microsatellite Instability for cancer detection and familial predisposition: development of international criteria for the determination of microsatellite instability in colorectal cancer.* In: 15 November 1998, pp. 5248–5257.

Bost, B., de Vienne, D., Hospital, F., Moreau, L. & Dillmann, C. (2001). Genetic and nongenetic bases for the L-shaped distribution of quantitative trait loci effects. *Genetics*. 157 (4). pp. 1773–

1787.

Botteri, E., Iodice, S., Raimondi, S., Maisonneuve, P. & Lowenfels, A.B. (2008). Cigarette smoking and adenomatous polyps: a meta-analysis. *Gastroenterology*. 134 (2). pp. 388–395.

Boyer, J.C. & Farber, R.A. (1998). Mutation rate of a microsatellite sequence in normal human fibroblasts. *Cancer research*. 58 (17). pp. 3946–3949.

Bronner, C.E., Baker, S.M., Morrison, P.T., Warren, G., Smith L.G., Lescoe, M.K., Kane, M., Earabino, C., Lipford, J. & Lindblom, A. (1994). Mutation in the DNA mismatch repair gene homologue hMLH1 is associated with hereditary non-polyposis colorectal cancer. Nature. 368 (6468). pp. 258-261

Burn, J., Chapman, P., Delhanty, J., Wood, C., Lalloo, F., Cachon-Gonzalez, M.B., Tsioupra, K., Church, W., Rhodes, M. & Gunn, A. (1991). The UK Northern region genetic register for familial adenomatous polyposis coli: use of age of onset, congenital hypertrophy of the retinal pigment epithelium, and DNA markers in risk calculations. *Journal of medical genetics*. 28 (5). pp. 289–296.

Cahill, D.P., Lengauer, C., Yu, J., Riggins, G.J., Willson, J.K., Markowitz, S.D., Kinzler, K.W. & Vogelstein, B. (1998). Mutations of mitotic checkpoint genes in human cancers. *Nature*. 392 (6673). pp. 300–303.

Cancer Genome Atlas Network (2012). Comprehensive molecular characterization of human colon and rectal cancer. *Nature*. 487 (7407). pp. 330–337.

Cannon-Albright, L.A., Goldgar, D.E., Meyer, L.J., Lewis, C.M., Anderson, D.E., Fountain, J.W., Hegi, M.E., Wiseman, R.W., Petty, E.M. & Bale, A.E. (1992). Assignment of a locus for familial melanoma, MLM, to chromosome 9p13-p22. *Science.* 258 (5085). pp. 1148–1152.

Casiano C.A., Mediavilla-Varela M., Tan E.M. (2006). Tumor-associated antigen arrays for the serological diagnosis of cancer. *Molecular & Cellular Proteomics.* 5 (10). pp.1745-59.

Cazier, J.-B. & Tomlinson, I. (2010). General lessons from large-scale studies to identify human cancer predisposition genes. *The Journal of pathology*. 220 (2). pp. 255–262.

Center, M.M., Jemal, A., Smith, R.A. & Ward, E. (2009). Worldwide variations in colorectal cancer. *CA: a cancer journal for clinicians*. 59 (6). pp. 366–378.

Chapman, P.B., Hauschild, A., Robert, C., Haanen, J.B., Ascierto, P., Larkin, J., Dummer, R., Garbe, C., Testori, A., Maio, M., Hogg, D., Lorigan, P., Lebbe, C., Jouary, T., Schadendorf, D., Ribas, A., O'Day, S.J., Sosman, J.A., Kirkwood, J.M., Eggermont, A.M.M., Dreno, B., Nolop, K., Li, J., Nelson, B., Hou, J., Lee, R.J., Flaherty, K.T., McArthur, G.A.BRIM-3 Study Group (2011). Improved survival with vemurafenib in melanoma with BRAF V600E mutation. *The New England journal of medicine*. 364 (26). pp. 2507–2516.

Chia, W.K., Ali, R. & Toh, H.C. (2012). Aspirin as adjuvant therapy for colorectal cancer--reinterpreting paradigms. *Nature reviews. Clinical oncology*. 9 (10). pp. 561–570.

Chipuk, J.E., Kuwana, T., Bouchier-Hayes, L., Droin, N.M., Newmeyer, D.D., Schuler, M. & Green, D.R. (2004). Direct activation of Bax by p53 mediates mitochondrial membrane permeabilization and apoptosis. *Science.* 303 (5660). pp. 1010–1014.

Compton, C.C. & Greene, F.L. (2004). The staging of colorectal cancer: 2004 and beyond. *CA: a cancer journal for clinicians*. 54 (6). pp. 295–308.

Conti, E. & Izaurralde, E. (2005). Nonsense-mediated mRNA decay: molecular insights and mechanistic variations across species. *Current opinion in cell biology*. 17 (3). pp. 316–325.

Croce, C.M. (2008). Oncogenes and cancer. *The New England journal of medicine*. 358 (5). pp. 502–511.

Croce, C.M. & Calin, G.A. (2005). miRNAs, cancer, and stem cell division. *Cell*. 122 (1). pp. 6–7.

Derycke, M.S., Gunawardena, S.R., Middha, S., Asmann, Y.W., Schaid, D.J., McDonnell, S.K., Riska, S.M., Eckloff, B.W., Cunningham, J.M., Fridley, B.L., Serie, D.J., Bamlet, W.R., Cicek, M.S., Jenkins, M.A., Duggan, D.J., Buchanan, D., Clendenning, M., Haile, R.W., Woods, M.O., Gallinger, S.N., Casey, G., Potter, J.D., Newcomb, P.A., Le Marchand, L., Lindor, N.M., Thibodeau, S.N. & Goode, E.L. (2013). Identification of Novel Variants in Colorectal Cancer Families by High-Throughput Exome Sequencing. *Cancer epidemiology, biomarkers & prevention*. 22 (7). pp. 1239–1251.

Downward, J. (2003). Targeting RAS signalling pathways in cancer therapy. *Nature reviews. Cancer*. 3 (1). pp. 11–22.

Dunican, D.S., McWilliam, P., Tighe, O., Parle-McDermott, A. & Croke, D.T. (2002). Gene expression differences between the microsatellite instability (MIN) and chromosomal instability (CIN) phenotypes in colorectal cancer revealed by high-density cDNA array hybridization. *Oncogene*. 21 (20). pp. 3253–3257.

Dunlop, M.G., Dobbins, S.E., Farrington, S.M., Jones, A.M., Palles, C., Whiffin, N., Tenesa, A., Spain, S., Broderick, P., Ooi, L.-Y., Domingo, E., Smillie, C., Henrion, M., Frampton, M., Martin, L., Grimes, G., Gorman, M., Semple, C., Ma, Y.P., Barclay, E., Prendergast, J., Cazier, J.-B., Olver, B., Penegar, S., Lubbe, S., Chander, I., Carvajal-Carmona, L.G., Ballereau, S., Lloyd, A., Vijayakrishnan, J., Zgaga, L., Rudan, I., Theodoratou, E., Colorectal Tumour Gene Identification (CORGI) Consortium, Starr, J.M., Deary, I., Kirac, I., Kovacević, D., Aaltonen, L.A., Renkonen-Sinisalo, L., Mecklin, J.-P., Matsuda, K., Nakamura, Y., Okada, Y., Gallinger, S., Duggan, D.J., Conti, D., Newcomb, P., Hopper, J., Jenkins, M.A., Schumacher, F., Casey, G., Easton, D., Shah, M., Pharoah, P., Lindblom, A., Liu, T., Swedish Low-Risk Colorectal Cancer Study Group, Smith, C.G., West, H., Cheadle, J.P., COIN Collaborative Group, Midgley, R., Kerr, D.J., Campbell, H., Tomlinson, I.P. & Houlston, R.S. (2012). Common variation near CDKN1A, POLD3 and SHROOM2 influences colorectal cancer risk. *Nature genetics*. 44 (7). pp. 770–776.

Dunlop, M.G., Tenesa, A., Farrington, S.M., Ballereau, S., Brewster, D.H., Koessler, T., Pharoah, P., Schafmayer, C., Hampe, J., Völzke, H., Chang-Claude, J., Hoffmeister, M., Brenner, H., Holst, von, S., Picelli, S., Lindblom, A., Jenkins, M.A., Hopper, J.L., Casey, G., Duggan, D., Newcomb, P.A., Abulí, A., Bessa, X., Ruiz-Ponte, C., Castellví-Bel, S., Niittymäki, I., Tuupanen, S., Karhu, A., Aaltonen, L., Zanke, B., Hudson, T., Gallinger, S., Barclay, E., Martin, L., Gorman, M., Carvajal-Carmona, L., Walther, A., Kerr, D., Lubbe, S., Broderick, P., Chandler, I., Pittman, A., Penegar, S., Campbell, H., Tomlinson, I. & Houlston, R.S. (2013). Cumulative impact of common genetic variants and other risk factors on colorectal cancer risk in 42,103 individuals. *Gut*. 62 (6). pp. 871–881.

Dyson, J.K. & Rutter, M.D. (2012). Colorectal cancer in inflammatory bowel disease: what is the real magnitude of the risk? *World journal of gastroenterology*. 18 (29). pp. 3839–3848.

El-Bchiri, J., Buhard, O., Penard-Lacronique, V., Thomas, G., Hamelin, R. & Duval, A. (2005). Differential nonsense mediated decay of mutated mRNAs in mismatch repair deficient colorectal cancers. *Human molecular genetics*. 14 (16). pp. 2435–2442.

Ellegren, H. (2004). Microsatellites: simple sequences with complex evolution. *Nature reviews. Genetics*. 5 (6). pp. 435–445.

Fearon, E.R. & Vogelstein, B. (1990). A genetic model for colorectal tumorigenesis. *Cell*. 61 (5). pp. 759–767.

Felix, R.S., Colleoni, G.W.B., Caballero, O.L., Yamamoto, M., Almeida, M.S.S., Andrade, V.C.C., Chauffaille, M. de L.L.F., Silva, W.A.D., Begnami, M.D., Soares, F.A., Simpson, A.J., Zago, M.A. & Vettore, A.L. (2009). SAGE analysis highlights the importance of p53csv, ddx5, mapkapk2 and ranbp2 to multiple myeloma tumorigenesis. *Cancer letters*. 278 (1). pp. 41–48.

Fevr, T., Robine, S., Louvard, D. & Huelsken, J. (2007). Wnt/beta-catenin is essential for intestinal homeostasis and maintenance of intestinal stem cells. *Molecular and cellular biology*. 27 (21). pp. 7551–7559.

Fishel, R., Lescoe, M.K., Rao, M.R., Copeland, N.G., Jenkins, N.A., Garber, J., Kane, M. & Kolodner, R. (1993). The human mutator gene homolog MSH2 and its association with hereditary nonpolyposis colon cancer. *Cell*. 75 (5). pp. 1027–1038.

Fisk, H.A., Mattison, C.P. & Winey, M. (2003). Human Mps1 protein kinase is required for centrosome duplication and normal mitotic progression. *Proceedings of the National Academy of Sciences of the United States of America*. 100 (25). pp. 14875–14880.

Fletcher, O. & Houlston, R.S. (2010). Architecture of inherited susceptibility to common cancer. *Nature reviews. Cancer*. 10 (5). pp. 353–361.

Fodde, R., Smits, R. & Clevers, H. (2001). APC, signal transduction and genetic instability in colorectal cancer. *Nature reviews. Cancer*. 1 (1). pp. 55–67.

Friedl, W., Meuschel, S., Caspari, R., Lamberti, C., Krieger, S., Sengteller, M. & Propping, P. (1996). Attenuated familial adenomatous polyposis due to a mutation in the 3' part of the APC gene. A clue for understanding the function of the APC protein. *Human genetics*. 97 (5). pp. 579–584.

Friend, S.H., Bernards, R., Rogelj, S., Weinberg, R.A., Rapaport, J.M., Albert, D.M. & Dryja, T.P. (1986). A human DNA segment with properties of the gene that predisposes to retinoblastoma and osteosarcoma. *Nature*. 323 (6089). pp. 643–646.

Futreal, P.A., Coin, L., Marshall, M., Down, T., Hubbard, T., Wooster, R., Rahman, N. & Stratton, M.R. (2004). A census of human cancer genes. *Nature reviews. Cancer*. 4 (3). pp. 177–183.

Galiatsatos, P. & Foulkes, W.D. (2006). Familial adenomatous polyposis. *The American journal of gastroenterology*. 101 (2). pp. 385–398.

Gibson, G. (2011). Rare and common variants: twenty arguments. *Nature reviews. Genetics*. 13 (2). pp. 135–145.

Giovannucci, E. (2002). Obesity, gender, and colon cancer. *Gut*. 51 (2). p. 147.

Goldgar, D.E., Easton, D.F., Cannon-Albright, L.A. & Skolnick, M.H. (1994). Systematic population-based assessment of cancer risk in first-degree relatives of cancer probands. *Journal of the National Cancer Institute*. 86 (21). pp. 1600–1608.

Grady, W.M. (2004). Genomic instability and colon cancer. *Cancer metastasis reviews*. 23 (1-2). pp. 11–27.

Grant, R.C., Al-Sukhni, W., Borgida, A.E., Holter, S., Kanji, Z.S., McPherson, T., Whelan, E., Serra, S., Trinh, Q.M., Peltekova, V., Stein, L.D., McPherson, J.D. & Gallinger, S. (2013). Exome sequencing identifies nonsegregating nonsense ATM and PALB2 variants in familial pancreatic cancer. *Human genomics*. 7. p. 11.

Gravel, S., Henn, B.M., Gutenkunst, R.N., Indap, A.R., Marth, G.T., Clark, A.G., Yu, F., Gibbs, R.A., 1000 Genomes Project & Bustamante, C.D. (2011). Demographic history and rare allele sharing among human populations. *Proceedings of the National Academy of Sciences of the United States of America*. 108 (29). pp. 11983–11988.

Greenman, C., Stephens, P., Smith, R., Dalgliesh, G.L., Hunter, C., Bignell, G., Davies, H., Teague, J., Butler, A., Stevens, C., Edkins, S., O'Meara, S., Vastrik, I., Schmidt, E.E., Avis, T., Barthorpe, S., Bhamra, G., Buck, G., Choudhury, B., Clements, J., Cole, J., Dicks, E., Forbes, S., Gray, K., Halliday, K., Harrison, R., Hills, K., Hinton, J., Jenkinson, A., Jones, D., Menzies, A., Mironenko, T., Perry, J., Raine, K., Richardson, D., Shepherd, R., Small, A., Tofts, C., Varian, J., Webb, T., West, S., Widaa, S., Yates, A., Cahill, D.P., Louis, D.N., Goldstraw, P., Nicholson, A.G., Brasseur, F., Looijenga, L., Weber, B.L., Chiew, Y.-E., DeFazio, A., Greaves, M.F., Green, A.R., Campbell, P., Birney, E., Easton, D.F., Chenevix-Trench, G., Tan, M.-H., Khoo, S.K., Teh, B.T., Yuen, S.T., Leung, S.Y., Wooster, R., Futreal, P.A. & Stratton, M.R. (2007). Patterns of somatic mutation in human cancer genomes. *Nature*. 446 (7132). pp. 153–158.

Groden J., Thliveris A., Samowitz W., Carlson M., Gelbert L., Albertsen H., Joslyn G., Stevens J., Spirio L., Robertson M. (1991) Identification and characterization of the familial adenomatous polyposis coli gene. *Cell*. 66 (3). pp.589-600.

Gudmundsson, J., Sulem, P., Gudbjartsson, D.F., Masson, G., Agnarsson, B.A., Benediktsdottir, K.R., Sigurdsson, A., Magnusson, O.T., Gudjonsson, S.A., Magnusdottir, D.N., Johannsdottir, H., Helgadottir, H.T., Stacey, S.N., Jonasdottir, A., Olafsdottir, S.B., Thorleifsson, G., Jonasson, J.G., Tryggvadottir, L., Navarrete, S., Fuertes, F., Helfand, B.T., Hu, Q., Csiki, I.E., Mates, I.N., Jinga, V., Aben, K.K.H., van Oort, I.M., Vermeulen, S.H., Donovan, J.L., Hamdy, F.C., Ng, C.-F., Chiu, P.K.F., Lau, K.-M., Ng, M.C.Y., Gulcher, J.R., Kong, A., Catalona, W.J., Mayordomo, J.I., Einarsson, G.V., Barkardottir, R.B., Jonsson, E., Mates, D., Neal, D.E., Kiemeney, L.A., Thorsteinsdottir, U., Rafnar, T. & Stefansson, K. (2012). A study based on whole-genome sequencing yields a rare variant at 8q24 associated with prostate cancer. *Nature genetics*. 44 (12). pp. 1326–1329.

Hall, J.M., Lee, M.K., Newman, B., Morrow, J.E., Anderson, L.A., Huey, B. & King, M.C. (1990). Linkage of early-onset familial breast cancer to chromosome 17q21. *Science*. 250 (4988). pp. 1684–1689.

Hanahan, D. & Weinberg, R.A. (2011). Hallmarks of cancer: the next generation. *Cell*. 144 (5). pp. 646–674.

Hanahan, D. & Weinberg, R.A. (2000). The hallmarks of cancer. *Cell*. 100 (1). pp. 57–70.

Hausen, zur, H. & de Villiers, E.M. (1994). Human papillomaviruses. *Annual review of microbiology*. 48. pp. 427–447.

He, T.C., Sparks, A.B., Rago, C., Hermeking, H., Zawel, L., da Costa, L.T., Morin, P.J., Vogelstein, B. & Kinzler, K.W. (1998). Identification of c-MYC as a target of the APC pathway. *Science*. 281 (5382). pp. 1509–1512.

Heinemann V., Douillard J.Y., Ducreux M., Peeters M.(2013).Targeted therapy in metastatic colorectal cancer -- an example of personalised medicine in action. *Cancer Treatment Reviews.* 39 (6). pp.592-601.

Hemminki, A., Markie, D., Tomlinson, I., Avizienyte, E., Roth, S., Loukola, A., Bignell, G., Warren, W., Aminoff, M., Höglund, P., Järvinen, H., Kristo, P., Pelin, K., Ridanpää, M., Salovaara, R., Toro, T., Bodmer, W., Olschwang, S., Olsen, A.S., Stratton, M.R., la Chapelle, de, A. & Aaltonen, L.A. (1998). A serine/threonine kinase gene defective in Peutz-Jeghers syndrome. *Nature*. 391 (6663). pp. 184–187.

Hemminki, A., Peltomäki, P., Mecklin, J.P., Järvinen, H., Salovaara, R., Nyström-Lahti, M., la Chapelle, de, A. & Aaltonen, L.A. (1994). Loss of the wild type MLH1 gene is a feature of hereditary nonpolyposis colorectal cancer. *Nature genetics*. 8 (4). pp. 405–410.

Hemminki A, Tomlinson I, Markie D, Järvinen H, Sistonen P, Björkqvist AM, Knuutila S, Salovaara R, Bodmer W, Shibata D, de la Chapelle A, Aaltonen LA. (1997). Localization of a susceptibility locus for Peutz-Jeghers syndrome to 19p using comparative genomic hybridization and targeted linkage analysis. *Nature Genetics*. 15 (1). pp. 87-90.

Houlston, R.S., Webb, E., Broderick, P., Pittman, A.M., Di Bernardo, M.C., Lubbe, S., Chandler, I., Vijayakrishnan, J., Sullivan, K., Penegar, S., Colorectal Cancer Association Study Consortium, Carvajal-Carmona, L., Howarth, K., Jaeger, E., Spain, S.L., Walther, A., Barclay, E., Martin, L., Gorman, M., Domingo, E., Teixeira, A.S., CORGI Consortium, Kerr, D., Cazier, J.-B., Niittymäki, I., Tuupanen, S., Karhu, A., Aaltonen, L.A., Tomlinson, I.P.M., Farrington, S.M., Tenesa, A., Prendergast, J.G.D., Barnetson, R.A., Cetnarskyj, R., Porteous, M.E., Pharoah, P.D.P., Koessler, T., Hampe, J., Buch, S., Schafmayer, C., Tepel, J., Schreiber, S., Völzke, H., Chang-Claude, J., Hoffmeister, M., Brenner, H., Zanke, B.W., Montpetit, A., Hudson, T.J., Gallinger, S., International Colorectal Cancer Genetic Association Consortium, Campbell, H. & Dunlop, M.G. (2008). Meta-analysis of genome-wide association data identifies four new susceptibility loci for colorectal cancer. *Nature genetics*. 40 (12). pp. 1426–1435.

Howe, J.R., Bair, J.L., Sayed, M.G., Anderson, M.E., Mitros, F.A., Petersen, G.M., Velculescu, V.E., Traverso, G. & Vogelstein, B. (2001). Germline mutations of the gene encoding bone morphogenetic protein receptor 1A in juvenile polyposis. *Nature genetics*. 28 (2). pp. 184–187.

Howe J.R., Ringold J.C., Summers R.W., Mitros F.A., Nishimura D.Y., Stone E.M. (1998a). A gene for familial juvenile polyposis maps to chromosome 18q21.1. American Journal of Human Genetics. 62. pp. 1129-1136.

Howe, J.R., Roth, S., Ringold, J.C., Summers, R.W., Järvinen, H.J., Sistonen, P., Tomlinson, I.P., Houlston, R.S., Bevan, S., Mitros, F.A., Stone, E.M. & Aaltonen, L.A. (1998b). Mutations in the SMAD4/DPC4 gene in juvenile polyposis. *Science*. 280 (5366). pp. 1086–1088.

Hsieh, P.-C., Chang, J.-C., Sun, W.-T., Hsieh, S.-C., Wang, M.-C. & Wang, F.-F. (2007). p53 downstream target DDA3 is a novel microtubule-associated protein that interacts with end-binding protein EB3 and activates beta-catenin pathway. *Oncogene*. 26 (34). pp. 4928–4940.

Hsieh, S.-C., Lo, P.-K. & Wang, F.-F. (2002). Mouse DDA3 gene is a direct transcriptional target of p53 and p73. *Oncogene*. 21 (19). pp. 3050–3057.

Humphries, A. & Wright, N.A. (2008). Colonic crypt organization and tumorigenesis. *Nature reviews. Cancer*. 8 (6). pp. 415–424.

Ikenoue, T., Kanai, F., Hikiba, Y., Obata, T., Tanaka, Y., Imamura, J., Ohta, M., Jazag, A., Guleng, B., Tateishi, K., Asaoka, Y., Matsumura, M., Kawabe, T. & Omata, M. (2005). Functional analysis of PIK3CA gene mutations in human colorectal cancer. *Cancer research*. 65 (11). pp. 4562–4567.

Ionov, Y., Peinado, M.A., Malkhosyan, S., Shibata, D. & Perucho, M. (1993). Ubiquitous somatic

mutations in simple repeated sequences reveal a new mechanism for colonic carcinogenesis. *Nature*. 363 (6429). pp. 558–561.

Ionov, Y., Yamamoto, H., Krajewski, S., Reed, J.C. & Perucho, M. (2000). Mutational inactivation of the proapoptotic gene BAX confers selective advantage during tumor clonal evolution. *Proceedings of the National Academy of Sciences of the United States of America*. 97 (20). pp. 10872–10877.

Isken, O. & Maquat, L.E. (2007). Quality control of eukaryotic mRNA: safeguarding cells from abnormal mRNA function. *Genes & development*. 21 (15). pp. 1833–1856.

Jakkula, E., Rehnström, K., Varilo, T., Pietiläinen, O.P.H., Paunio, T., Pedersen, N.L., deFaire, U., Järvelin, M.-R., Saharinen, J., Freimer, N., Ripatti, S., Purcell, S., Collins, A., Daly, M.J., Palotie, A. & Peltonen, L. (2008). The genome-wide patterns of variation expose significant substructure in a founder population. *American journal of human genetics*. 83 (6). pp. 787–794.

Jass, J.R. (2004). HNPCC and sporadic MSI-H colorectal cancer: a review of the morphological similarities and differences. *Familial cancer*. 3 (2). pp. 93–100.

Joslyn, G., Carlson, M., Thliveris, A. Albertsen, H., Gelbert, L., Samowitz, W., Groden, J., Stevens, J., Spirio, L. , Roberston, M. (1991). Identification of deletion mutations and three new genes at the familial polyposis locus. *Cell*. 9;66(3). pp. 601-13.

Järvinen, H.J. (1992). Epidemiology of familial adenomatous polyposis in Finland: impact of family screening on the colorectal cancer rate and survival. *Gut*. 33 (3). pp. 357–360.

Järvinen, H.J. (2003). Genetic testing for polyposis: practical and ethical aspects. *Gut*. 52 Suppl 2. pp. ii19–22.

Jemal, A., Siegel, R., Ward, E., Hao, Y., Xu, J., Murray, T. & Thun, M.J. (2008). Cancer statistics, 2008. *CA: a cancer journal for clinicians*. 58 (2). pp. 71–96.

Jen, J., Powell, S.M., Papadopoulos, N., Smith, K.J., Hamilton, S.R., Vogelstein, B. & Kinzler, K.W. (1994). Molecular determinants of dysplasia in colorectal lesions. *Cancer research*. 54 (21). pp. 5523–5526.

Jenkins, M.A., Croitoru, M.E., Monga, N., Cleary, S.P., Cotterchio, M., Hopper, J.L. & Gallinger, S. (2006). Risk of colorectal cancer in monoallelic and biallelic carriers of MYH mutations: a population-based case-family study. *Cancer epidemiology, biomarkers & prevention.* 15 (2). pp. 312–314.

Jeon, B.-N., Choi, W.-I., Yu, M.-Y., Yoon, A.-R., Kim, M.-H., Yun, C.-O. & Hur, M.-W. (2009). ZBTB2, a novel master regulator of the p53 pathway. *The Journal of biological chemistry*. 284 (27). pp. 17935–17946.

Jiricny, J. (2006). The multifaceted mismatch-repair system. *Nature reviews. Molecular cell biology*. 7 (5). pp. 335–346.

Jo, W.-S. & Chung, D.C. (2005). Genetics of hereditary colorectal cancer. *Seminars in oncology*. 32 (1). pp. 11–23.

Johns, L.E. & Houlston, R.S. (2001). A systematic review and meta-analysis of familial colorectal cancer risk. *The American journal of gastroenterology*. 96 (10). pp. 2992–3003.

Jones, S., Hruban, R.H., Kamiyama, M., Borges, M., Zhang, X., Parsons, D.W., Lin, J.C.-H.,

Palmisano, E., Brune, K., Jaffee, E.M., Iacobuzio-Donahue, C.A., Maitra, A., Parmigiani, G., Kern, S.E., Velculescu, V.E., Kinzler, K.W., Vogelstein, B., Eshleman, J.R., Goggins, M. & Klein, A.P. (2009). Exomic sequencing identifies PALB2 as a pancreatic cancer susceptibility gene. *Science*. 324 (5924). p. 217.

Kane, M.F., Loda, M., Gaida, G.M., Lipman, J., Mishra, R., Goldman, H., Jessup, J.M. & Kolodner, R. (1997). Methylation of the hMLH1 promoter correlates with lack of expression of hMLH1 in sporadic colon tumors and mismatch repair-defective human tumor cell lines. *Cancer research*. 57 (5). pp. 808–811.

Kiezun, A., Garimella, K., Do, R., Stitziel, N.O., Neale, B.M., McLaren, P.J., Gupta, N., Sklar, P., Sullivan, P.F., Moran, J.L., Hultman, C.M., Lichtenstein, P., Magnusson, P., Lehner, T., Shugart, Y.Y., Price, A.L., de Bakker, P.I.W., Purcell, S.M. & Sunyaev, S.R. (2012). Exome sequencing and the genetic basis of complex traits. *Nature genetics*. 44 (6). pp. 623–630.

Kilpivaara, O. & Aaltonen, L.A. (2013). Diagnostic cancer genome sequencing and the contribution of germline variants. *Science*. 339 (6127). pp. 1559–1562.

Kim, W.K., Park, M., Park, M., Kim, Y.J., Shin, N., Kim, H.K., You, K.T. & Kim, H. (2013). Identification and selective degradation of neopeptide-containing truncated mutant proteins in the tumors with high microsatellite instability. *Clinical cancer research*. 19 (13). pp. 3369–3382.

Kinzler K.W., Nilbert M.C., Su L.K., Vogelstein B., Bryan T.M., Levy D.B., Smith K.J., Preisinger A.C., Hedge P., McKechnie D., Finnier R., Matrkam A., Groffen J., Boguski M.S., Altschul S.F., Horii A., Ando H., Miyoshi Y., Miki Y., Nishisho I., Nakamura Y. (1991). Identification of FAP locus genes from chromosome 5q21. *Science*. 253 (5020). pp. 661-5

Kinzler, K.W. & Vogelstein, B. (1997). Cancer-susceptibility genes. Gatekeepers and caretakers. *Nature*. 386 (6627). pp. 761–763.

Kinzler, K.W. & Vogelstein, B. (1996). Lessons from hereditary colorectal cancer. *Cell*. 87 (2). pp. 159–170.

Kirkwood, J.M., Butterfield, L.H.,Tarhini, A.A., Zarour, H., Kalinski, P., Ferrone, S. (2012). *CA: A Cancer Journal for Clinicians*. 62 (5). pp. 309-335.

Knudson, A.G. (2002). Cancer genetics. *American journal of medical genetics*. 111 (1). pp. 96–102.

Knudson, A.G. (1971). Mutation and cancer: statistical study of retinoblastoma. *Proceedings of the National Academy of Sciences of the United States of America*. 68 (4). pp. 820–823.

Knudson, A.G. (2001). Two genetic hits (more or less) to cancer. *Nature reviews. Cancer*. 1 (2). pp. 157–162.

Koepp, D.M., Schaefer, L.K., Ye, X., Keyomarsi, K., Chu, C., Harper, J.W. & Elledge, S.J. (2001). Phosphorylation-dependent ubiquitination of cyclin E by the SCFFbw7 ubiquitin ligase. *Science*. 294 (5540). pp. 173–177.

Kolodner, R.D., Hall, N.R., Lipford, J., Kane, M.F., Morrison, P.T., Finan, P.J., Burn, J., Chapman, P., Earabino, C. & Merchant, E. (1995). Structure of the human MLH1 locus and analysis of a large hereditary nonpolyposis colorectal carcinoma kindred for mlh1 mutations. *Cancer research*. 55 (2). pp. 242–248.

Kops, G.J.P.L., Weaver, B.A.A. & Cleveland, D.W. (2005). On the road to cancer: aneuploidy and the mitotic checkpoint. *Nature reviews. Cancer*. 5 (10). pp. 773–785.

la Chapelle, de, A. (2004). Genetic predisposition to colorectal cancer. *Nature reviews. Cancer*. 4 (10). pp. 769–780.

Lahiri, D.K. & Nurnberger, J.I. (1991). A rapid non-enzymatic method for the preparation of HMW DNA from blood for RFLP studies. *Nucleic acids research*. 19 (19). p. 5444.

Laken, S.J., Petersen, G.M., Gruber, S.B., Oddoux, C., Ostrer, H., Giardiello, F.M., Hamilton, S.R., Hampel, H., Markowitz, A., Klimstra, D., Jhanwar, S., Winawer, S., Offit, K., Luce, M.C., Kinzler, K.W. & Vogelstein, B. (1997). Familial colorectal cancer in Ashkenazim due to a hypermutable tract in APC. *Nature genetics*. 17 (1). pp. 79–83.

Lamlum, H., Ilyas, M., Rowan, A., Clark, S., Johnson, V., Bell, J., Frayling, I., Efstathiou, J., Pack, K., Payne, S., Roylance, R., Gorman, P., Sheer, D., Neale, K., Phillips, R., Talbot, I., Bodmer, W. & Tomlinson, I. (1999). The type of somatic mutation at APC in familial adenomatous polyposis is determined by the site of the germline mutation: a new facet to Knudson's "two-hit" hypothesis. *Nature medicine*. 5 (9). pp. 1071–1075.

Lamlum, H., Papadopoulou, A., Ilyas, M., Rowan, A., Gillet, C., Hanby, A., Talbot, I., Bodmer, W. & Tomlinson, I. (2000). APC mutations are sufficient for the growth of early colorectal adenomas. *Proceedings of the National Academy of Sciences of the United States of America*. 97 (5). pp. 2225–2228.

Lammi, L., Arte, S., Somer, M., Järvinen, H., Lahermo, P., Thesleff, I., Pirinen, S. & Nieminen, P. (2004). Mutations in AXIN2 cause familial tooth agenesis and predispose to colorectal cancer. *American journal of human genetics*. 74 (5). pp. 1043–1050.

Leach, F.S., Nicolaides, N.C., Papadopoulos, N., Liu, B., Jen, J., Parsons, R., Peltomäki, P., Sistonen, P., Aaltonen, L.A. & Nyström-Lahti, M. (1993). Mutations of a mutS homolog in hereditary nonpolyposis colorectal cancer. *Cell*. 75 (6). pp. 1215–1225.

Leary, R.J., Lin, J.C., Cummins, J., Boca, S., Wood, L.D., Parsons, D.W., Jones, S., Sjöblom, T., Park, B.H., Parsons, R., Willis, J., Dawson, D., Willson, J.K.V., Nikolskaya, T., Nikolsky, Y., Kopelovich, L., Papadopoulos, N., Pennacchio, L.A., Wang, T.-L., Markowitz, S.D., Parmigiani, G., Kinzler, K.W., Vogelstein, B. & Velculescu, V.E. (2008). Integrated analysis of homozygous deletions, focal amplifications, and sequence alterations in breast and colorectal cancers. *Proceedings of the National Academy of Sciences of the United States of America*. 105 (42). pp. 16224–16229.

Lee, I.-M., Shiroma, E.J., Lobelo, F., Puska, P., Blair, S.N., Katzmarzyk, P.T.Lancet Physical Activity Series Working Group (2012). Effect of physical inactivity on major non-communicable diseases worldwide: an analysis of burden of disease and life expectancy. *Lancet*. 380 (9838). pp. 219–229.

Leng M., Chan D.W., Luo H., Zhu C., Qin J., Wang Y. (2006). MPS1-dependent mitotic BLM phosphorylation is important for chromosome stability. *Proceedings of the National Academy of Sciences of the United States of America*. 103 (31). pp.11485-90

Levy, D.B., Smith, K.J., Beazer-Barclay, Y., Hamilton, S.R., Vogelstein, B. & Kinzler, K.W. (1994). Inactivation of both APC alleles in human and mouse tumors. *Cancer research*. 54 (22). pp. 5953–5958.

Ley, T.J., Mardis, E.R., Ding, L., Fulton, B., McLellan, M.D., Chen, K., Dooling, D., Dunford-Shore, B.H., McGrath, S., Hickenbotham, M., Cook, L., Abbott, R., Larson, D.E., Koboldt, D.C., Pohl, C., Smith, S., Hawkins, A., Abbott, S., Locke, D., Hillier, L.W., Miner, T., Fulton, L., Magrini, V., Wylie, T., Glasscock, J., Conyers, J., Sander, N., Shi, X., Osborne, J.R., Minx, P., Gordon,

D., Chinwalla, A., Zhao, Y., Ries, R.E., Payton, J.E., Westervelt, P., Tomasson, M.H., Watson, M., Baty, J., Ivanovich, J., Heath, S., Shannon, W.D., Nagarajan, R., Walter, M.J., Link, D.C., Graubert, T.A., DiPersio, J.F. & Wilson, R.K. (2008). DNA sequencing of a cytogenetically normal acute myeloid leukaemia genome. *Nature*. 456 (7218). pp. 66–72.

Li, H. & Durbin, R. (2009). Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics*. 25 (14). pp. 1754–1760.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R.1000 Genome Project Data Processing Subgroup (2009). The Sequence Alignment/Map format and SAMtools. *Bioinformatics*. 25 (16). pp. 2078–2079.

Lichtenstein, P., Holm, N.V., Verkasalo, P.K., Iliadou, A., Kaprio, J., Koskenvuo, M., Pukkala, E., Skytthe, A. & Hemminki, K. (2000). Environmental and heritable factors in the causation of cancer--analyses of cohorts of twins from Sweden, Denmark, and Finland. *The New England journal of medicine*. 343 (2). pp. 78–85.

Lin, Z., Gao, C., Ning, Y., He, X., Wu, W. & Chen, Y.-G. (2008). The pseudoreceptor BMP and activin membrane-bound inhibitor positively modulates Wnt/beta-catenin signaling. *The Journal of biological chemistry*. 283 (48). pp. 33053–33058.

Lindblom, A., Tannergård, P., Werelius, B. & Nordenskjöld, M. (1993). Genetic mapping of a second locus predisposing to hereditary non-polyposis colon cancer. *Nature genetics*. 5 (3). pp. 279–282.

Liu, S.-T., Chan, G.K.T., Hittle, J.C., Fujii, G., Lees, E. & Yen, T.J. (2003). Human MPS1 kinase is required for mitotic arrest induced by the loss of CENP-E from kinetochores. *Molecular biology of the cell*. 14 (4). pp. 1638–1651.

Liu J, Cheng X, Zhang Y, Li S, Cui H, Zhang L, Shi R, Zhao Z, He C, Wang C, Zhao H, Zhang C, Fisk HA, Guadagno TM, Cui Y. (2013). Phosphorylation of Mps1 by BRAFV600E prevents Mps1 degradation and contributes to chromosome instability in melanoma. *Oncogene*. 32 (6). pp.713-23

Loeb, L.A. (1991). Mutator phenotype may be required for multistage carcinogenesis. *Cancer research*. 51 (12). pp. 3075–3079.

Loeb, L.A., Loeb, K.R. & Anderson, J.P. (2003). Multiple mutations and cancer. *Proceedings of the National Academy of Sciences of the United States of America*. 100 (3). pp. 776–781.

Lubbe, S.J., Di Bernardo, M.C., Broderick, P., Chandler, I. & Houlston, R.S. (2012). Comprehensive evaluation of the impact of 14 genetic variants on colorectal cancer phenotype and risk. *American journal of epidemiology*. 175 (1). pp. 1–10.

Lubbe, S.J., Webb, E.L., Chandler, I.P. & Houlston, R.S. (2009). Implications of familial colorectal cancer risk profiles and microsatellite instability status. *Journal of clinical oncology*. 27 (13). pp. 2238–2244.

Luo, L., Shen, G.-Q., Stiffler, K.A., Wang, Q.K., Pretlow, T.G. & Pretlow, T.P. (2006). Loss of heterozygosity in human aberrant crypt foci (ACF), a putative precursor of colon cancer. *Carcinogenesis*. 27 (6). pp. 1153–1159.

Lynch, H.T. & Lynch, P.M. (2005). Molecular screening for the Lynch syndrome--better than family history? *The New England journal of medicine*. 352 (18). pp. 1920–1922.

Lynch, H.T., Boland, C.R., Gong, G., Shaw, T.G., Lynch, P.M., Fodde, R., Lynch, J.F. & la Chapelle,

de, A. (2006). Phenotypic and genotypic heterogeneity in the Lynch syndrome: diagnostic, surveillance and management implications. *European journal of human genetics : EJHG*. 14 (4). pp. 390–402.

Lynch, H.T., Smyrk, T., McGinn, T., Lanspa, S., Cavalieri, J., Lynch, J., Slominski-Castor, S., Cayouette, M.C., Priluck, I. & Luce, M.C. (1995). Attenuated familial adenomatous polyposis (AFAP). A phenotypically and genotypically distinctive variant of FAP. *Cancer*. 76 (12). pp. 2427–2433.

Lynch, P.M. (1999). Clinical challenges in management of familial adenomatous polyposis and hereditary nonpolyposis colorectal cancer. *Cancer*. 86 (11 Suppl). pp. 2533–2539.

MacArthur, D.G., Balasubramanian, S., Frankish, A., Huang, N., Morris, J., Walter, K., Jostins, L., Habegger, L., Pickrell, J.K., Montgomery, S.B., Albers, C.A., Zhang, Z.D., Conrad, D.F., Lunter, G., Zheng, H., Ayub, Q., DePristo, M.A., Banks, E., Hu, M., Handsaker, R.E., Rosenfeld, J.A., Fromer, M., Jin, M., Mu, X.J., Khurana, E., Ye, K., Kay, M., Saunders, G.I., Suner, M.-M., Hunt, T., Barnes, I.H.A., Amid, C., Carvalho-Silva, D.R., Bignell, A.H., Snow, C., Yngvadottir, B., Bumpstead, S., Cooper, D.N., Xue, Y., Romero, I.G., 1000 Genomes Project Consortium, Wang, J., Li, Y., Gibbs, R.A., McCarroll, S.A., Dermitzakis, E.T., Pritchard, J.K., Barrett, J.C., Harrow, J., Hurles, M.E., Gerstein, M.B. & Tyler-Smith, C. (2012). A systematic survey of loss-of-function variants in human protein-coding genes. *Science*. 335 (6070). pp. 823–828.

Maciejowski, J., George, K.A., Terret, M.-E., Zhang, C., Shokat, K.M. & Jallepalli, P.V. (2010). Mps1 directs the assembly of Cdc20 inhibitory complexes during interphase and mitosis to control M phase timing and spindle checkpoint signaling. *The Journal of cell biology*. 190 (1). pp. 89–100.

Mackay, T.F. (2001). The genetic architecture of quantitative traits. *Annual review of genetics*. 35. pp. 303–339.

Manolio, T.A., Collins, F.S., Cox, N.J., Goldstein, D.B., Hindorff, L.A., Hunter, D.J., McCarthy, M.I., Ramos, E.M., Cardon, L.R., Chakravarti, A., Cho, J.H., Guttmacher, A.E., Kong, A., Kruglyak, L., Mardis, E., Rotimi, C.N., Slatkin, M., Valle, D., Whittemore, A.S., Boehnke, M., Clark, A.G., Eichler, E.E., Gibson, G., Haines, J.L., Mackay, T.F.C., McCarroll, S.A. & Visscher, P.M. (2009). Finding the missing heritability of complex diseases. *Nature*. 461 (7265). pp. 747–753.

Mao, M., Tian, F., Mariadason, J.M., Tsao, C.C., Lemos, R., Dayyani, F., Gopal, Y.N.V., Jiang, Z.-Q., Wistuba, I.I., Tang, X.M., Bornman, W.G., Bollag, G., Mills, G.B., Powis, G., Desai, J., Gallick, G.E., Davies, M.A. & Kopetz, S. (2013). Resistance to BRAF inhibition in BRAF-mutant colon cancer can be overcome with PI3K inhibition or demethylating agents. *Clinical cancer research*. 19 (3). pp. 657–667.

Markowitz, S., Wang, J., Myeroff, L., Parsons, R., Sun, L., Lutterbaugh, J., Fan, R.S., Zborowska, E., Kinzler, K.W. & Vogelstein, B. (1995). Inactivation of the type II TGF-beta receptor in colon cancer cells with microsatellite instability. *Science*. 268 (5215). pp. 1336–1338.

Markowitz, S.D. & Bertagnolli, M.M. (2009). Molecular origins of cancer: Molecular basis of colorectal cancer. *The New England journal of medicine*. 361 (25). pp. 2449–2460.

Massagué, J. (2008). TGFbeta in Cancer. *Cell*. 134 (2). pp. 215–230.

Maurano, M.T., Humbert, R., Rynes, E., Thurman, R.E., Haugen, E., Wang, H., Reynolds, A.P., Sandstrom, R., Qu, H., Brody, J., Shafer, A., Neri, F., Lee, K., Kutyavin, T., Stehling-Sun, S., Johnson, A.K., Canfield, T.K., Giste, E., Diegel, M., Bates, D., Hansen, R.S., Neph, S., Sabo, P.J., Heimfeld, S., Raubitschek, A., Ziegler, S., Cotsapas, C., Sotoodehnia, N., Glass, I., Sunyaev,

S.R., Kaul, R. & Stamatoyannopoulos, J.A. (2012). Systematic localization of common disease-associated variation in regulatory DNA. *Science*. 337 (6099). pp. 1190–1195.

McCarthy, M.I., Abecasis, G.R., Cardon, L.R., Goldstein, D.B., Little, J., Ioannidis, J.P.A. & Hirschhorn, J.N. (2008). Genome-wide association studies for complex traits: consensus, uncertainty and challenges. *Nature reviews. Genetics*. 9 (5). pp. 356–369.

McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., Garimella, K., Altshuler, D., Gabriel, S., Daly, M. & DePristo, M.A. (2010). The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome research*. 20 (9). pp. 1297–1303.

McLaren, W., Pritchard, B., Rios, D., Chen, Y., Flicek, P. & Cunningham, F. (2010). Deriving the consequences of genomic variants with the Ensembl API and SNP Effect Predictor. *Bioinformatics*. 26 (16). pp. 2069–2070.

Mecklin, J.-P. & Järvinen, H.J. (2005). Surveillance in Lynch syndrome. *Familial cancer*. 4 (3). pp. 267–271.

Mehlen, P. & Puisieux, A. (2006). Metastasis: a question of life or death. *Nature reviews. Cancer*. 6 (6). pp. 449–458.

Meijers-Heijboer, H., van den Ouweland, A., Klijn, J., Wasielewski, M., de Snoo, A., Oldenburg, R., Hollestelle, A., Houben, M., Crepin, E., van Veghel-Plandsoen, M., Elstrodt, F., van Duijn, C., Bartels, C., Meijers, C., Schutte, M., McGuffog, L., Thompson, D., Easton, D., Sodha, N., Seal, S., Barfoot, R., Mangion, J., Chang-Claude, J., Eccles, D., Eeles, R., Evans, D.G., Houlston, R., Murday, V., Narod, S., Peretz, T., Peto, J., Phelan, C., Zhang, H.X., Szabo, C., Devilee, P., Goldgar, D., Futreal, P.A., Nathanson, K.L., Weber, B., Rahman, N., Stratton, M.R.CHEK2-Breast Cancer Consortium (2002). Low-penetrance susceptibility to breast cancer due to CHEK2(*)1100delC in noncarriers of BRCA1 or BRCA2 mutations. *Nature genetics*. 31 (1). pp. 55–59.

Merlo, L.M.F., Pepper, J.W., Reid, B.J. & Maley, C.C. (2006). Cancer as an evolutionary and ecological process. *Nature reviews. Cancer*. 6 (12). pp. 924–935.

Michor, F., Iwasa, Y., Vogelstein, B., Lengauer, C. & Nowak, M.A. (2005). Can chromosomal instability initiate tumorigenesis? *Seminars in cancer biology*. 15 (1). pp. 43–49.

Miki, Y., Swensen, J., Shattuck-Eidens, D., Futreal, P.A., Harshman, K., Tavtigian, S., Liu, Q., Cochran, C., Bennett, L.M. & Ding, W. (1994). A strong candidate for the breast and ovarian cancer susceptibility gene BRCA1. *Science*. 266 (5182). pp. 66–71.

Misale, S., Yaeger, R., Hobor, S., Scala, E., Janakiraman, M., Liska, D., Valtorta, E., Schiavo, R., Buscarino, M., Siravegna, G., Bencardino, K., Cercek, A., Chen, C.-T., Veronese, S., Zanon, C., Sartore-Bianchi, A., Gambacorta, M., Gallicchio, M., Vakiani, E., Boscaro, V., Medico, E., Weiser, M., Siena, S., Di Nicolantonio, F., Solit, D. & Bardelli, A. (2012). Emergence of KRAS mutations and acquired resistance to anti-EGFR therapy in colorectal cancer. *Nature*. 486 (7404). pp. 532–536.

Miyaki, M., Konishi, M., Tanaka, K., Kikuchi-Yanoshita, R., Muraoka, M., Yasuno, M., Igari, T., Koike, M., Chiba, M. & Mori, T. (1997). Germline mutation of MSH6 as the cause of hereditary nonpolyposis colorectal cancer. *Nature genetics*. 17 (3). pp. 271–272.

Miyashita, T. & Reed, J.C. (1995). Tumor suppressor p53 is a direct transcriptional activator of the human bax gene. *Cell*. 80 (2). pp. 293–299.

Mori, Y., Sato, F., Selaru, F.M., Olaru, A., Perry, K., Kimos, M.C., Tamura, G., Matsubara, N., Wang, S., Xu, Y., Yin, J., Zou, T.-T., Leggett, B., Young, J., Nukiwa, T., Stine, O.C., Abraham, J.M., Shibata, D. & Meltzer, S.J. (2002). Instabilotyping reveals unique mutational spectra in microsatellite-unstable gastric cancers. *Cancer research*. 62 (13). pp. 3641–3645.

Morin, P.J., Sparks, A.B., Korinek, V., Barker, N., Clevers, H., Vogelstein, B. & Kinzler, K.W. (1997). Activation of beta-catenin-Tcf signaling in colon cancer by mutations in beta-catenin or APC. *Science*. 275 (5307). pp. 1787–1790.

Nagy, E. & Maquat, L.E. (1998). A rule for termination-codon position within intron-containing genes: when nonsense affects RNA abundance. *Trends in biochemical sciences*. 23 (6). pp. 198–199.

Nagy, R., Sweet, K. & Eng, C. (2004). Highly penetrant hereditary cancer syndromes. *Oncogene*. 23 (38). pp. 6445–6470.

Navarro, M.S. & Bachant, J. (2008). RanBP2: a tumor suppressor with a new twist on TopoII, SUMO, and centromeres. *Cancer cell*. 13 (4). pp. 293–295.

Nick McElhinny, S.A., Gordenin, D.A., Stith, C.M., Burgers, P.M.J. & Kunkel, T.A. (2008). Division of labor at the eukaryotic replication fork. *Molecular cell*. 30 (2). pp. 137–144.

Nicolaides, N.C., Papadopoulos, N., Liu, B., Wei, Y.F., Carter, K.C., Ruben, S.M., Rosen, C.A., Haseltine, W.A., Fleischmann, R.D. & Fraser, C.M. (1994). Mutations of two PMS homologues in hereditary nonpolyposis colon cancer. *Nature*. 371 (6492). pp. 75–80.

Niittymäki, I., Kaasinen, E., Tuupanen, S., Karhu, A., Järvinen, H., Mecklin, J.-P., Tomlinson, I.P.M., Di Bernardo, M.C., Houlston, R.S. & Aaltonen, L.A. (2010). Low-penetrance susceptibility variants in familial colorectal cancer. *Cancer epidemiology, biomarkers & prevention*. 19 (6). pp. 1478–1483.

Nishisho, I., Nakamura, Y., Miyoshi, Y., Miki, Y., Ando, H., Horii, A., Koyama, K., Utsunomiya, J., Baba, S. & Hedge, P. (1991). Mutations of chromosome 5q21 genes in FAP and colorectal cancer patients. *Science*. 253 (5020). pp. 665–669.

Nowak, M.A., Komarova, N.L., Sengupta, A., Jallepalli, P.V., Shih, I.-M., Vogelstein, B. & Lengauer, C. (2002). The role of chromosomal instability in tumor initiation. *Proceedings of the National Academy of Sciences of the United States of America*. 99 (25). pp. 16226–16231.

Nyström-Lahti, M., Kristo, P., Nicolaides, N.C., Chang, S.Y., Aaltonen, L.A., Moisio, A.L., Järvinen, H.J., Mecklin, J.P., Kinzler, K.W. & Vogelstein, B. (1995). Founding mutations and Alu-mediated recombination in hereditary colon cancer. *Nature medicine*. 1 (11). pp. 1203–1206.

Nyström-Lahti, M., Wu, Y., Moisio, A.L., Hofstra, R.M., Osinga, J., Mecklin, J.P., Järvinen, H.J., Leisti, J., Buys, C.H., la Chapelle, de, A. & Peltomäki, P. (1996). DNA mismatch repair gene mutations in 55 kindreds with verified or putative hereditary non-polyposis colorectal cancer. *Human molecular genetics*. 5 (6). pp. 763–769.

Palles, C., Cazier, J.-B., Howarth, K.M., Domingo, E., Jones, A.M., Broderick, P., Kemp, Z., Spain, S.L., Guarino, E., Guarino Almeida, E., Salguero, I., Sherborne, A., Chubb, D., Carvajal-Carmona, L.G., Ma, Y., Kaur, K., Dobbins, S., Barclay, E., Gorman, M., Martin, L., Kovac, M.B., Humphray, S., CORGI Consortium, WGS500 Consortium, Lucassen, A., Holmes, C.C., Bentley, D., Donnelly, P., Taylor, J., Petridis, C., Roylance, R., Sawyer, E.J., Kerr, D.J., Clark, S., Grimes, J., Kearsey, S.E., Thomas, H.J.W., McVean, G., Houlston, R.S. & Tomlinson, I. (2013). Germline mutations affecting the proofreading domains of POLE and POLD1 predispose

to colorectal adenomas and carcinomas. *Nature genetics*. 45 (2). pp. 136–144.

Papadopoulos, N., Nicolaides, N.C., Wei, Y.F., Ruben, S.M., Carter, K.C., Rosen, C.A., Haseltine, W.A., Fleischmann, R.D., Fraser, C.M. & Adams, M.D. (1994). Mutation of a mutL homolog in hereditary colon cancer. *Science*. 263 (5153). pp. 1625-1629.

Parsonnet, J., Friedman, G.D., Vandersteen, D.P., Chang, Y., Vogelman, J.H., Orentreich, N. & Sibley, R.K. (1991). Helicobacter pylori infection and the risk of gastric carcinoma. *The New England journal of medicine*. 325 (16). pp. 1127–1131.

Parsons, R., Li, G.M., Longley, M.J., Fang, W.H., Papadopoulos, N., Jen, J., la Chapelle, de, A., Kinzler, K.W., Vogelstein, B. & Modrich, P. (1993). Hypermutability and mismatch repair deficiency in RER+ tumor cells. *Cell*. 75 (6). pp. 1227–1236.

Peltomäki, P. (2005). Lynch syndrome genes. *Familial cancer*. 4 (3). pp. 227–232.

Peltomäki, P. & Vasen, H. (2004). Mutations associated with HNPCC predisposition -- Update of ICG-HNPCC/INSiGHT mutation database. *Disease markers*. 20 (4-5). pp. 269–276.

Peltomäki, P., Aaltonen, L.A., Sistonen, P., Pylkkänen, L., Mecklin, J.P., Järvinen, H., Green, J.S., Jass, J.R., Weber, J.L. & Leach, F.S. (1993). Genetic mapping of a locus predisposing to human colorectal cancer. *Science*. 260 (5109). pp. 810–812.

Peltonen, L., Palotie, A. & Lange, K. (2000). Use of population isolates for mapping complex traits. *Nature reviews. Genetics*. 1 (3). pp. 182–190.

Piepkorn, M. (2000). Melanoma genetics: an update with focus on the CDKN2A(p16)/ARF tumor suppressors. *Journal of the American Academy of Dermatology*. 42 (5 Pt 1). pp. 705–22– quiz 723–6.

Polakis, P. (2002). Casein kinase 1: a Wnt'er of disconnect. *Current biology : CB*. 12 (14). pp. R499–R501.

Polakis, P. (2007). The many ways of Wnt in cancer. *Current opinion in genetics & development*. 17 (1). pp. 45–51.

Pomerantz, M.M., Ahmadiyeh, N., Jia, L., Herman, P., Verzi, M.P., Doddapaneni, H., Beckwith, C.A., Chan, J.A., Hills, A., Davis, M., Yao, K., Kehoe, S.M., Lenz, H.-J., Haiman, C.A., Yan, C., Henderson, B.E., Frenkel, B., Barretina, J., Bass, A., Tabernero, J., Baselga, J., Regan, M.M., Manak, J.R., Shivdasani, R., Coetzee, G.A. & Freedman, M.L. (2009). The 8q24 cancer risk variant rs6983267 shows long-range interaction with MYC in colorectal cancer. *Nature genetics*. 41 (8). pp. 882–884.

Powell, S.M., Zilz, N., Beazer-Barclay, Y., Bryan, T.M., Hamilton, S.R., Thibodeau, S.N., Vogelstein, B. & Kinzler, K.W. (1992). APC mutations occur early during colorectal tumorigenesis. *Nature*. 359 (6392). pp. 235–237.

Prives, C. (1998). Signaling to p53: breaking the MDM2-p53 circuit. *Cell*. 95 (1). pp. 5–8.

Pylayeva-Gupta, Y., Grabocka, E. & Bar-Sagi, D. (2011). RAS oncogenes: weaving a tumorigenic web. *Nature reviews. Cancer*. 11 (11). pp. 761–774.

Rajagopalan, H., Bardelli, A., Lengauer, C., Kinzler, K.W., Vogelstein, B. & Velculescu, V.E. (2002). Tumorigenesis: RAF/RAS oncogenes and mismatch-repair status. *Nature*. 418 (6901). p. 934.

Rajagopalan, H., Nowak, M.A., Vogelstein, B. & Lengauer, C. (2003). The significance of unstable chromosomes in colorectal cancer. *Nature reviews. Cancer*. 3 (9). pp. 695–701.

Rampino, N., Yamamoto, H., Ionov, Y., Li, Y., Sawai, H., Reed, J.C. & Perucho, M. (1997). Somatic frameshift mutations in the BAX gene in colon cancers of the microsatellite mutator phenotype. *Science*. 275 (5302). pp. 967–969.

Reddy, E.P., Reynolds, R.K., Santos, E. & Barbacid, M. (1982). A point mutation is responsible for the acquisition of transforming properties by the T24 human bladder carcinoma oncogene. *Nature*. 300 (5888). pp. 149–152.

Renner, O., Fominaya, J., Alonso, S., Blanco-Aparicio, C., Leal, J.F.M. & Carnero, A. (2007). Mst1, RanBP2 and eIF4G are new markers for in vivo PI3K activation in murine and human prostate. *Carcinogenesis*. 28 (7). pp. 1418–1425.

Roberts, N.J., Jiao, Y., Yu, J., Kopelovich, L., Petersen, G.M., Bondy, M.L., Gallinger, S., Schwartz, A.G., Syngal, S., Cote, M.L., Axilbund, J., Schulick, R., Ali, S.Z., Eshleman, J.R., Velculescu, V.E., Goggins, M., Vogelstein, B., Papadopoulos, N., Hruban, R.H., Kinzler, K.W. & Klein, A.P. (2012). ATM mutations in patients with hereditary pancreatic cancer. *Cancer discovery*. 2 (1). pp. 41–46.

Rodrigues, N.R., Rowan, A., Smith, M.E., Kerr, I.B., Bodmer, W.F., Gannon, J.V. & Lane, D.P. (1990). p53 mutations in colorectal cancer. *Proceedings of the National Academy of Sciences of the United States of America*. 87 (19). pp. 7555–7559.

Rozen, P., Samuel, Z., Shomrat, R. & Legum, C. (1999). Notable intrafamilial phenotypic variability in a kindred with familial adenomatous polyposis and an APC mutation in exon 9. *Gut*. 45 (6). pp. 829–833.

Salovaara, R., Loukola, A., Kristo, P., Kääriäinen, H., Ahtola, H., Eskelinen, M., Härkönen, N., Julkunen, R., Kangas, E., Ojala, S., Tulikoura, J., Valkamo, E., Järvinen, H., Mecklin, J.P., Aaltonen, L.A. & la Chapelle, de, A. (2000). Population-based molecular detection of hereditary nonpolyposis colorectal cancer. *Journal of clinical oncology*. 18 (11). pp. 2193–2200.

Sammalkorpi, H., Alhopuro, P., Lehtonen, R., Tuimala, J., Mecklin, J.-P., Järvinen, H.J., Jiricny, J., Karhu, A. & Aaltonen, L.A. (2007). Background mutation frequency in microsatellite-unstable colorectal cancer. *Cancer research*. 67 (12). pp. 5691–5698.

Samuels, Y. & Ericson, K. (2006). Oncogenic PI3K and its role in cancer. *Current opinion in oncology*. 18 (1). pp. 77–82.

Samuels, Y. & Velculescu, V.E. (2004). Oncogenic mutations of PIK3CA in human cancers. *Cell cycle*. 3 (10). pp. 1221–1224.

Santarosa, M. & Ashworth, A. (2004). Haploinsufficiency for tumour suppressor genes: when you don't need to go all the way. *Biochimica et biophysica acta*. 1654 (2). pp. 105–122.

Schreibman, I.R., Baker, M., Amos, C. & McGarrity, T.J. (2005). The hamartomatous polyposis syndromes: a clinical and molecular review. *The American journal of gastroenterology*. 100 (2). pp. 476–490.

Segditsas, S., Rowan, A.J., Howarth, K., Jones, A., Leedham, S., Wright, N.A., Gorman, P., Chambers, W., Domingo, E., Roylance, R.R., Sawyer, E.J., Sieber, O.M. & Tomlinson, I.P.M. (2009). APC and the three-hit hypothesis. *Oncogene*. 28 (1). pp. 146–155.

Seshagiri, S., Stawiski, E.W., Durinck, S., Modrusan, Z., Storm, E.E., Conboy, C.B., Chaudhuri, S., Guan, Y., Janakiraman, V., Jaiswal, B.S., Guillory, J., Ha, C., Dijkgraaf, G.J.P., Stinson, J., Gnad, F., Huntley, M.A., Degenhardt, J.D., Haverty, P.M., Bourgon, R., Wang, W., Koeppen, H., Gentleman, R., Starr, T.K., Zhang, Z., Largaespada, D.A., Wu, T.D. & de Sauvage, F.J. (2012). Recurrent R-spondin fusions in colon cancer. *Nature*. 488 (7413). pp. 660–664.

Sharma, S.V., Bell, D.W., Settleman, J., Haber, D.A. (2007). Epidermal growth factor receptor mutations in lung cancer. *Nature Reviews Cancer.* 7. pp. 169-181.

Shen, H. & Laird, P.W. (2013). Interplay between the cancer genome and epigenome. *Cell*. 153 (1). pp. 38–55.

Shendure, J. & Ji, H. (2008). Next-generation DNA sequencing. *Nature biotechnology*. 26 (10). pp. 1135–1145.

Shih, I.M., Zhou, W., Goodman, S.N., Lengauer, C., Kinzler, K.W. & Vogelstein, B. (2001). Evidence that genetic instability occurs at an early stage of colorectal tumorigenesis. *Cancer research*. 61 (3). pp. 818–822.

Shitashige, M., Satow, R., Honda, K., Ono, M., Hirohashi, S. & Yamada, T. (2008). Regulation of Wnt signaling by the nuclear pore complex. *Gastroenterology*. 134 (7). pp. 1961–71– 1971.e1–4.

Shitoh, K., Furukawa, T., Kojima, M., Konishi, F., Miyaki, M., Tsukamoto, T. & Nagai, H. (2001). Frequent activation of the beta-catenin-Tcf signaling pathway in nonfamilial colorectal carcinomas with microsatellite instability. *Genes, chromosomes & cancer*. 30 (1). pp. 32–37.

Sieber O.M., Lipton L., Crabtree M., Heinimann K., Fidalgo P., Phillips R.K., Bisgaard M.L., Orntoft T.F., Aaltonen L.A., Hodgson S.V., Thomas H.J., Tomlinson I.P.(2003). Multiple colorectal adenomas, classic adenomatous polyposis, and germ-line mutations in MYH.*New Englend Journal of Medicine*. 348 (9) pp. 791-9.

Sjöblom, T., Jones, S., Wood, L.D., Parsons, D.W., Lin, J., Barber, T.D., Mandelker, D., Leary, R.J., Ptak, J., Silliman, N., Szabo, S., Buckhaults, P., Farrell, C., Meeh, P., Markowitz, S.D., Willis, J., Dawson, D., Willson, J.K.V., Gazdar, A.F., Hartigan, J., Wu, L., Liu, C., Parmigiani, G., Park, B.H., Bachman, K.E., Papadopoulos, N., Vogelstein, B., Kinzler, K.W. & Velculescu, V.E. (2006). The consensus coding sequences of human breast and colorectal cancers. *Science*. 314 (5797). pp. 268–274.

Sliedrecht, T., Zhang, C., Shokat, K.M. & Kops, G.J.P.L. (2010). Chemical genetic inhibition of Mps1 in stable human cell lines reveals novel aspects of Mps1 function in mitosis. *PloS one*. 5 (4). p. e10251.

Smith, C.G., Naven, M., Harris, R., Colley, J., West, H., Li, N., Liu, Y., Adams, R., Maughan, T.S., Nichols, L., Kaplan, R., Wagner, M.J., McLeod, H.L. & Cheadle, J.P. (2013). Exome resequencing identifies potential tumor-suppressor genes that predispose to colorectal cancer. *Human mutation*. 34 (7). pp. 1026–1034.

Smith, S.A., Easton, D.F., Evans, D.G. & Ponder, B.A. (1992). Allele losses in the region 17q12-21 in familial breast and ovarian cancer involve the wild-type chromosome. *Nature genetics*. 2 (2). pp. 128–131.

Stoffel, E., Mukherjee, B., Raymond, V.M., Tayob, N., Kastrinos, F., Sparr, J., Wang, F., Bandipalliam, P., Syngal, S. & Gruber, S.B. (2009). Calculation of risk of colorectal and endometrial cancer among patients with Lynch syndrome. *Gastroenterology*. 137 (5). pp. 1621–1627.

Stratton, M.R., Campbell, P.J. & Futreal, P.A. (2009). The cancer genome. *Nature*. 458 (7239). pp. 719–724.

Sullivan, P.F., Daly, M.J. & O'Donovan, M. (2012). Genetic architectures of psychiatric disorders: the emerging picture and its implications. *Nature reviews. Genetics*. 13 (8). pp. 537–551.

Sun, T., Yang, X., Wang, W., Zhang, X., Xu, Q., Zhu, S., Kuchta, R., Chen, G. & Liu, X. (2010). Cellular abundance of Mps1 and the role of its carboxyl terminal tail in substrate recruitment. *The Journal of biological chemistry*. 285 (49). pp. 38730–38739.

Sun, W.-T., Hsieh, P.-C., Chiang, M.-L., Wang, M.-C. & Wang, F.-F. (2008). p53 target DDA3 binds ASPP2 and inhibits its stimulation on p53-mediated BAX activation. *Biochemical and biophysical research communications*. 376 (2). pp. 395–398.

Sur, I.K., Hallikas, O., Vähärautio, A., Yan, J., Turunen, M., Enge, M., Taipale, M., Karhu, A., Aaltonen, L.A. & Taipale, J. (2012). Mice lacking a Myc enhancer that includes human SNP rs6983267 are resistant to intestinal tumors. *Science*. 338 (6112). pp. 1360–1363.

Swanton, C., Marani, M., Pardo, O., Warne, P.H., Kelly, G., Sahai, E., Elustondo, F., Chang, J., Temple, J., Ahmed, A.A., Brenton, J.D., Downward, J. & Nicke, B. (2007). Regulators of mitotic arrest and ceramide metabolism are determinants of sensitivity to paclitaxel and other chemotherapeutic drugs. *Cancer cell*. 11 (6). pp. 498–512.

Tannergård, P., Zabarovsky, E., Stanbridge, E., Nordenskjöld, M. & Lindblom, A. (1994). Sublocalization of a locus at 3p21.3-23 predisposing to hereditary nonpolyposis colon cancer. *Human genetics*. 94 (2). pp. 210–214.

Tetsu, O. & McCormick, F. (1999). Beta-catenin regulates expression of cyclin D1 in colon carcinoma cells. *Nature*. 398 (6726). pp. 422–426.

Thibodeau, S.N., Bren, G. & Schaid, D. (1993). Microsatellite instability in cancer of the proximal colon. *Science*. 260 (5109). pp. 816–819.

Tomlinson, I., Webb, E., Carvajal-Carmona, L., Broderick, P., Kemp, Z., Spain, S., Penegar, S., Chandler, I., Gorman, M., Wood, W., Barclay, E., Lubbe, S., Martin, L., Sellick, G., Jaeger, E., Hubner, R., Wild, R., Rowan, A., Fielding, S., Howarth, K., CORGI Consortium, Silver, A., Atkin, W., Muir, K., Logan, R., Kerr, D., Johnstone, E., Sieber, O., Gray, R., Thomas, H., Peto, J., Cazier, J.-B. & Houlston, R. (2007). A genome-wide association scan of tag SNPs identifies a susceptibility variant for colorectal cancer at 8q24.21. *Nature genetics*. 39 (8). pp. 984–988.

Tuupanen, S., Turunen, M., Lehtonen, R., Hallikas, O., Vanharanta, S., Kivioja, T., Björklund, M., Wei, G., Yan, J., Niittymäki, I., Mecklin, J.-P., Järvinen, H., Ristimäki, A., Di-Bernardo, M., East, P., Carvajal-Carmona, L., Houlston, R.S., Tomlinson, I., Palin, K., Ukkonen, E., Karhu, A., Taipale, J. & Aaltonen, L.A. (2009). The common colorectal cancer predisposition SNP rs6983267 at chromosome 8q24 confers potential to enhanced Wnt signaling. *Nature genetics*. 41 (8). pp. 885–890.

Umar, A., Boland, C.R., Terdiman, J.P., Syngal, S., la Chapelle, de, A., Rüschoff, J., Fishel, R., Lindor, N.M., Burgart, L.J., Hamelin, R., Hamilton, S.R., Hiatt, R.A., Jass, J., Lindblom, A., Lynch, H.T., Peltomäki, P., Ramsey, S.D., Rodriguez-Bigas, M.A., Vasen, H.F.A., Hawk, E.T., Barrett, J.C., Freedman, A.N. & Srivastava, S. (2004). *Revised Bethesda Guidelines for hereditary nonpolyposis colorectal cancer (Lynch syndrome) and microsatellite instability.* In: 18 February 2004, pp. 261–268.

Vahteristo, P., Bartkova, J., Eerola, H., Syrjäkoski, K., Ojala, S., Kilpivaara, O., Tamminen, A.,

Kononen, J., Aittomäki, K., Heikkilä, P., Holli, K., Blomqvist, C., Bartek, J., Kallioniemi, O.-P. & Nevanlinna, H. (2002). A CHEK2 genetic variant contributing to a substantial fraction of familial breast cancer. *American journal of human genetics*. 71 (2). pp. 432–438.

Van Schaeybroeck, S., Allen, W.L., Turkington, R.C. & Johnston, P.G. (2011). Implementing prognostic and predictive biomarkers in CRC clinical trials. *Nature reviews. Clinical oncology*. 8 (4). pp. 222–232.

Varjosalo, M., Sacco, R., Stukalov, A., van Drogen, A., Planyavsky, M., Hauri, S., Aebersold, R., Bennett, K.L., Colinge, J., Gstaiger, M. & Superti-Furga, G. (2013). Interlaboratory reproducibility of large-scale human protein-complex analysis by standardized AP-MS. *Nature methods*. 10 (4). pp. 307–314.

Vasen, H.F., Mecklin, J.P., Khan, P.M. & Lynch, H.T. (1991). *The International Collaborative Group on Hereditary Non-Polyposis Colorectal Cancer (ICG-HNPCC).* In: May 1991, pp. 424–425.

Vasen, H.F., Watson, P., Mecklin, J.P. & Lynch, H.T. (1999). *New clinical criteria for hereditary nonpolyposis colorectal cancer (HNPCC, Lynch syndrome) proposed by the International Collaborative group on HNPCC.* In: June 1999, pp. 1453–1456.

Veigl, M.L., Kasturi, L., Olechnowicz, J., Ma, A.H., Lutterbaugh, J.D., Periyasamy, S., Li, G.M., Drummond, J., Modrich, P.L., Sedwick, W.D. & Markowitz, S.D. (1998). Biallelic inactivation of hMLH1 by epigenetic gene silencing, a novel mechanism causing human MSI cancers. *Proceedings of the National Academy of Sciences of the United States of America*. 95 (15). pp. 8698–8702.

Vetter, I.R., Nowak, C., Nishimoto, T., Kuhlmann, J. & Wittinghofer, A. (1999). Structure of a Ran-binding domain complexed with Ran bound to a GTP analogue: implications for nuclear transport. *Nature*. 398 (6722). pp. 39–46.

Vilar, E. & Gruber, S.B. (2010). Microsatellite instability in colorectal cancer-the stable evidence. *Nature reviews. Clinical oncology*. 7 (3). pp. 153–162.

Vogelstein, B. & Kinzler, K.W. (2004). Cancer genes and the pathways they control. *Nature medicine*. 10 (8). pp. 789–799.

Vogelstein, B., Fearon, E.R., Hamilton, S.R., Kern, S.E., Preisinger, A.C., Leppert, M., Nakamura, Y., White, R., Smits, A.M. & Bos, J.L. (1988). Genetic alterations during colorectal-tumor development. *The New England journal of medicine*. 319 (9). pp. 525–532.

Vogelstein, B., Lane, D. & Levine, A.J. (2000). Surfing the p53 network. *Nature*. 408 (6810)16 November. pp. 307–310.

Vogelstein, B., Papadopoulos, N., Velculescu, V.E., Zhou, S., Diaz, L.A. & Kinzler, K.W. (2013). Cancer genome landscapes. *Science (New York, N.Y.)*. 339 (6127). pp. 1546–1558.

Wan, P.T.C., Garnett, M.J., Roe, S.M., Lee, S., Niculescu-Duvaz, D., Good, V.M., Jones, C.M., Marshall, C.J., Springer, C.J., Barford, D., Marais, R.Cancer Genome Project (2004). Mechanism of activation of the RAF-ERK signaling pathway by oncogenic mutations of B-RAF. *Cell*. 116 (6). pp. 855–867.

Wang, J., Sun, L., Myeroff, L., Wang, X., Gentry, L.E., Yang, J., Liang, J., Zborowska, E., Markowitz, S. & Willson, J.K. (1995). Demonstration that mutation of the type II transforming growth factor beta receptor inactivates its tumor suppressor activity in replication error-positive colon carcinoma cells. *The Journal of biological chemistry*. 270 (37). pp. 22044–22049.

Wang, T., Wei, J.J., Sabatini, D.M., Lander, E. (2014) Genetic Screens in Human Cells Uing the CRISPR-Cas9 System. *Science.* 343 (6166). pp. 80-84

Watanabe, T., Wu, T.T., Catalano, P.J., Ueki, T., Satriano, R., Haller, D.G., Benson, A.B. & Hamilton, S.R. (2001). Molecular predictors of survival after adjuvant chemotherapy for colon cancer. *The New England journal of medicine*. 344 (16). pp. 1196–1206.

Weitz, J., Koch, M., Debus, J., Höhler, T., Galle, P.R., Buchler, M.W. (2005). Colorectal cancer. *The Lancet.* 365 (9454). pp. 153-165.

Welcker, M. & Clurman, B.E. (2008). FBW7 ubiquitin ligase: a tumour suppressor at the crossroads of cell division, growth and differentiation. *Nature reviews. Cancer*. 8 (2). pp. 83–93.

Wood, L.D., Parsons, D.W., Jones, S., Lin, J., Sjöblom, T., Leary, R.J., Shen, D., Boca, S.M., Barber, T., Ptak, J., Silliman, N., Szabo, S., Dezso, Z., Ustyanksky, V., Nikolskaya, T., Nikolsky, Y., Karchin, R., Wilson, P.A., Kaminker, J.S., Zhang, Z., Croshaw, R., Willis, J., Dawson, D., Shipitsin, M., Willson, J.K.V., Sukumar, S., Polyak, K., Park, B.H., Pethiyagoda, C.L., Pant, P.V.K., Ballinger, D.G., Sparks, A.B., Hartigan, J., Smith, D.R., Suh, E., Papadopoulos, N., Buckhaults, P., Markowitz, S.D., Parmigiani, G., Kinzler, K.W., Velculescu, V.E. & Vogelstein, B. (2007). The genomic landscapes of human breast and colorectal cancers. *Science*. 318 (5853). pp. 1108–1113.

Wooster, R., Bignell, G., Lancaster, J., Swift, S., Seal, S., Mangion, J., Collins, N., Gregory, S., Gumbs, C. & Micklem, G. (1995). Identification of the breast cancer susceptibility gene BRCA2. *Nature*. 378 (6559). pp. 789–792.

Xu, Y. & Pasche, B. (2007). TGF-beta signaling alterations and susceptibility to colorectal cancer. *Human molecular genetics*. 16 Spec No 1. pp. R14–20.

Yada, M., Hatakeyama, S., Kamura, T., Nishiyama, M., Tsunematsu, R., Imaki, H., Ishida, N., Okumura, F., Nakayama, K. & Nakayama, K.I. (2004). Phosphorylation-dependent degradation of c-Myc is mediated by the F-box protein Fbw7. *The EMBO journal*. 23 (10). pp. 2116–2125.

Zhu S., Wang W., Clarke D.C., Liu X. (2007) Activation of Mps1 promotes transforming growth factor-beta-independent Samd signaling. *The Journal of Biological Chemistry.*