MOLECULAR ALTERATIONS IN FAMILIAL COLORECTAL CANCER

Tuija Hienonen

Helsinki 2005
MOLECULAR ALTERATIONS IN FAMILIAL COLORECTAL CANCER

Tuija Hienonen

Department of Medical Genetics
Molecular and Cancer Biology Programme
Haartman Institute and Biomedicum Helsinki
University of Helsinki
Finland

Academic dissertation

To be publicly discussed, with the permission of the Medical Faculty of the University of Helsinki, in the lecture hall 2, Biomedicum Helsinki, Haartmaninkatu 8, on November 25th 2005, at 12 noon

Helsinki 2005
3.3.1 MLH3 sequencing and denaturing high-performance liquid chromatography (dHPLC) analysis .............................................................. 28
3.4 Genetic analyses of SEMG1 and intergenic T9 repeats (IV) ............... 28
  3.4.1 SEMG1 sequencing and denaturing high-performance liquid chromatography (dHPLC) analysis .............................................................. 28
  3.4.2 Evaluation of SEMG1 mRNA levels by quantitative PCR ............. 29
  3.4.3 Evaluation of SEMG1 protein levels by Western blotting .......... 29
  3.4.4 Sequencing of intergenic T9 repeats ............................................ 29

RESULTS .............................................................................................................. 30
  1. Common regions of allelic imbalance (I) ...................................................... 30
  2. Analyses of AURKA (II) ................................................................................. 30
  3. MLH3 mutation analysis (III) ....................................................................... 32
  4. Genetic analyses of SEMG1 and intergenic T9 repeats (IV) .................. 32

DISCUSSION ........................................................................................................ 34
  1. Novel colorectal cancer susceptibility loci (I) ............................................... 34
  2. AURKA, a novel low-penetrance susceptibility gene (II) ........................... 35
  3. Role of MLH3 in microsatellite-unstable colorectal cancer (III) ............. 37
  4. High frequency of microsatellite mutations in SEMG1 and intergenic T9 repeats (IV) ........................................................................................................ 37

CONCLUSIONS AND FUTURE PROSPECTS .................................................. 42

ACKNOWLEDGEMENTS .................................................................................. 43

REFERENCES ...................................................................................................... 45
ABBREVIATIONS

A  adenine  
ACF  aberrant crypt focus  
AFAP  attenuated familial adenomatous polyposis  
AI  allelic imbalance  
APC  adenomatous polyposis coli  
AURKA  aurora kinase A  
BAX  BLC2-associated X protein  
bp  base pair  
BER  base excision repair  
BRAF  v-raf murine sarcoma viral oncogene homolog B1  
BRCA2  breast cancer 2  
BUB1  budding uninhibited by benimidazoles 1 (yeast) homolog  
BUB1B  budding uninhibited by benimidazoles 1 (yeast) homolog beta  
C  cytosine  
CDK4  cyclin-dependent kinase 4  
CDKN2A  cyclin-dependent kinase inhibitor 2A  
cDNA  complementary deoxyribonucleic acid  
CGH  comparative genomic hybridization  
CIN  chromosomal instability  
cM  centiMorgan  
c-Myc  v-myc avian myelocytomatosis viral oncogene homolog  
CRC  colorectal cancer  
CS  Cowden syndrome  
Cyr61  cysteine-rich, angiogenic inducer, 61  
DCC  deleted in colorectal carcinoma  
dHPLC  denaturing high-performance liquid chromatography  
DNA  deoxyribonucleic acid  
DSBR  double strand break repair  
EST  expressed sequence tag  
EXO1  exonuclease 1  
FAP  familial adenomatous polyposis  
FMM  familial malignant melanoma  
G  guanine  
GAPDH  glyceraldehyde-3-phosphate dehydrogenase  
GI  gastrointestinal  
GIST  gastrointestinal stromal tumour  
GSTT1  glutathione S-transferase theta 1  
HNPCC  hereditary non-polyposis colorectal cancer  
HPRC  hereditary papillary renal carcinoma  
HRAS  Harvey rat sarcoma viral oncogene homolog  
IDL  insertion/deletion loop  
IGFII  insulin-like growth factor II  
JP  juvenile polyposis  
kb  kilobase  
KIT  v-kit Hardy-Zuckerman 4 feline sarcoma viral oncogene homolog
KRAS: Kirsten rat sarcoma viral oncogene homolog
LD: linkage disequilibrium
LEF: lymphoid enhancer factor, transcription factor
LKB1: serine/threonine kinase 11 (STK11)
LOH: loss of heterozygosity
LOI: loss of imprinting
MDM2: double minute 2 (mouse) homolog
MEN2: multiple endocrine neoplasia type 2
MET: met proto-oncogene (hepatocyte growth factor receptor)
MGMT: O⁶-methylguanine-DNA methyltransferase
Min: multiple intestinal neoplasia
MLH1, -3: MutL (E. coli) homologs 1 and 3
MMR: mismatch repair
Mom1, -2: modifier of multiple intestinal neoplasia 1 and 2
MRE11: meiotic recombination protein 11 (S. cerevisiae) homolog
mRNA: messenger ribonucleic acid
MSH2, -3, -6: MutS (E. coli) homologs 2, 3, and 6
MSI: microsatellite instability, microsatellite-unstable
MSI-H: high microsatellite instability
MSI-L: low microsatellite instability
MSS: microsatellite stability, microsatellite stable
MTHFR: 5,10-methylenetetrahydrofolate reductase
MTS: Muir-Torre syndrome
MYCL: v-myec avian myelocytomatosis viral oncogene homolog, lung carcinoma-derived
MYH: MutY (E. coli) homolog
NAT1, -2: N-acetyltransferases 1 and 2
NER: nucleotide excision repair
NF1: neurofibromin 1
NMD: nonsense-mediated decay
p: short arm of a chromosome
PCNA: proliferating cell nuclear antigen
PCR: polymerase chain reaction
PJS: Peutz-Jeghers syndrome
PLA2G2A: phospholipase A2, group IIA
PMS1, -2: postmeiotic segregation increased 1 and 2
PRL-3: protein tyrosine phosphatase type IVA, member 3
PTEN: phosphatase and tensin homolog
q: long arm of a chromosome
qPCR: quantitative polymerase chain reaction
QTL: quantitative trait locus
RB: retinoblastoma
RET: ret proto-oncogene
RNA: ribonucleic acid
SEMG1: semenogelin 1
SMAD2, -4: Mad (D. melanogaster) homologs 2 and 4
SNP: single nucleotide polymorphism
T: thymine
TCF: T-cell factor, transcription factor
TGF-β: transforming growth factor-β
<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>TGFβRII</td>
<td>transforming growth factor-β receptor type II</td>
</tr>
<tr>
<td>TP53</td>
<td>tumour protein p53</td>
</tr>
<tr>
<td>TS</td>
<td>Turcot syndrome</td>
</tr>
<tr>
<td>VNTR</td>
<td>variable number of tandem repeats</td>
</tr>
<tr>
<td>WNT</td>
<td>wingless-type</td>
</tr>
<tr>
<td>XP</td>
<td>xeroderma pigmentosum</td>
</tr>
<tr>
<td>ZNF217</td>
<td>zinc finger protein 217</td>
</tr>
</tbody>
</table>
LIST OF ORIGINAL PUBLICATIONS

This thesis is based on the following original articles referred to in the text by Roman numerals I-IV.


* Equal contribution

Publication I was included in the thesis of Päivi Laiho (Molecular classification of colorectal cancer, Helsinki 2005).

The original articles are reproduced with permission from the copyright holders.
ABSTRACT

The genetic aetiology of colorectal cancer (CRC) is one of the best characterized of all human cancer types. Genes involved in cell cycle regulation or DNA error repair systems have been shown to predispose to familial CRC. Most of the identified susceptibility alleles are rare in the general population and may thus explain only a minority of all CRC cases. However, the more prevalent but less well characterized low-penetrance susceptibility alleles could associate with a majority of heritable cases. More information on the specific low-penetrance susceptibility alleles is thus required.

CRC progresses via two main pathways: the chromosomal instability (CIN) and microsatellite instability (MSI) pathways. Most CRCs display CIN with gross chromosomal alterations but the exact mechanisms initiating the instability phenotype are not well characterized. A smaller subset of the CRCs display MSI, i.e. small deletions or insertions in short repetitive regions of the genome, as a result of defective DNA mismatch repair (MMR) system. This study aimed at identifying the molecular mechanisms involved in CRC predisposition and progression in both of the instability pathways.

Novel susceptibility loci were sought by identifying commonly deleted or amplified regions in familial CIN CRC cases. Especially chromosome 20q13 was seen to harbour amplifications more frequently in familial than in sporadic CRCs indicating the possible existence of novel predisposing oncogene(s) within that amplicon. A 20q13 gene AURKA was chosen for further analysis since it has been suggested to function as a low-penetrance susceptibility gene. A 91A-specific amplification pattern was observed in the familial cases. Furthermore, a trend between younger age at diagnosis and 91A was observed, supporting the role of this allele in CRC predisposition.

Around 15% of all CRCs display MSI and are associated with inactive MMR genes MLH1 or MSH2, and less frequently PMS2 or MSH6. Causative mutation has not been identified in a subset of the MSI cases, indicating the possible involvement of other genes. We investigated the putative role of MMR gene MLH3 in CRC predisposition. A patient with mild CRC clustering in the family harbouried a germline missense mutation that was not found in the cancer-free controls. The variant is not an attractive disease-causing candidate, thus MLH3 does not seem to play a major role in CRC predisposition.

We investigated the possible role of SEMG1 as a novel target for microsatellite mutations in MSI CRCs. Although frequently mutated, the SEMG1 microsatellite mutations do not seem to be selected for during tumorigenesis. Intergenic microsatellites were studied to obtain information on the general background mutation frequency of MSI CRCs. One microsatellite was frequently mutated, indicating that mutation frequency alone cannot be utilized to identify true MSI target genes.
REVIEW OF THE LITERATURE

1. Cancer genetics and epigenetics

Cancer originates from a single cell that acquires a clonal growth advantage over the surrounding cells. Additional genetic and epigenetic changes occurring in the progeny cells facilitate malignant transformation by producing genetically and biologically abnormal cells. This clonal progression model of cancer indicates that cancer formation requires several changes in a given cell (Nowell 1976). Thus age is a prominent factor in cancer susceptibility (Miller 1980) but cancer risk is higher in individuals with an inherited genetic predisposition.

Malignant transformation is a result of exogenous and endogenous factors disrupting cellular homeostasis normally controlled by the cell cycle, differentiation, and apoptosis pathways. It has been suggested that almost all cancer types display the following six features untypical of normal cells: self-sufficiency in growth signals, insensitivity to growth inhibitory signals, evasion of apoptosis, limitless replication potential, sustained angiogenesis, and metastasis (Hanahan and Weinberg 2000). These abnormal cellular features are caused by alterations in tumour suppressor genes and oncogenes that control the cellular homeostasis in an antagonist manner.

1.1 Tumour suppressor genes

Tumour suppressor genes inhibit abnormal cells from progressing in the cell cycle. Inactivation of both alleles of a tumour suppressor gene is required for uncontrolled cell growth, thus tumour suppressor genes are recessive at the cellular level (Knudson’s two-hit hypothesis) (Knudson 1971). Individuals with an inherited tumour suppressor gene mutation are at a higher cancer risk because all cells harbour one inactive allele. Inherited mutations in tumour suppressor genes are typically small changes such as point mutations or small insertions/deletions. The second alteration affecting the wild-type allele is often larger, such as deletion of a whole chromosome.

A haploinsufficiency model suggests that inactivation of one allele of certain tumour suppressor genes is enough to promote tumorigenesis. Evidence of some haploinsufficient tumour suppressor genes have been obtained, for example the colorectal cancer (CRC)-associated genes PTEN and LKB1 (Di Cristofano et al. 1998, Miyoshi et al. 2002, Rossi et al. 2002). Loss of the wild-type allele might provide a further growth advantage to the affected cells as suggested by the more frequent LKB1 loss of heterozygosity (LOH) observed in carcinomas than in polyps (Entius et al. 2001).

Tumour suppressor genes can be divided into three groups: gatekeepers, landcapers, and caretakers (Kinzler and Vogelstein 1997, Kinzler and Vogelstein 1998). Gatekeepers control cell growth by inhibiting cell cycle progression and promoting apoptosis of abnormal cells. Landscaper gene mutations affect the cellular microenvironment (stromal cells) rather than growth pattern of the neoplastic cells themselves. Caretaker tumour suppressor genes are involved in controlling genomic

1.1.1 DNA repair genes

DNA repair genes are caretaker tumour suppressor genes involved in controlling genomic stability. Exogenous and endogenous factors, such as radiation, reactive oxygen species, and DNA replication and recombination, constantly cause DNA sequence alterations that need to be corrected in order to maintain proper cellular functions. An array of DNA repair systems have been identified with at least five main DNA error correction mechanisms functioning in human cells: 1. base excision repair (BER) corrects residues modified by endogenous agents, 2. nucleotide excision repair (NER) restores mainly helix-distorting errors caused by environmental mutagens, 3. mismatch repair (MMR) corrects nucleotide mismatches and small insertion/deletion loops (IDLs) generated during DNA replication or recombination, 4. direct repair reverses DNA damage, and 5. double-strand break repair (DSBR) corrects breakage in the double helix (Hoeijmakers 2001, Friedberg 2003). Mutations in BER, NER, MMR, and DSBR genes have been shown to be causative in cancer predisposition syndromes, such as MYH-associated colorectal polyposis, xeroderma pigmentosum (XP), hereditary non-polyposis colorectal cancer (HNPCC), and breast cancer, respectively (Al-Tassan et al. 2002, Friedberg 2003).

1.2 Oncogenes

Oncogenes are excessively active counterparts of cellular proto-oncogenes functioning as, for example, growth factors, growth factor receptors, intracellular messengers, or transcription factors. Oncogenes promote cell proliferation by inducing mitosis and inhibiting differentiation and apoptosis. Gain of function mutations, such as point mutations, amplifications, or chromosomal rearrangements transform proto-oncogenes into excessively active oncogenes. Oncogenes function in a dominant manner, hence activating mutation in the other allele is sufficient to promote tumorigenesis (Bishop 1991).

Inherited oncogene mutations are rare, possibly because of negative selection pressure during embryonic development. Some known exceptions are germline mutations in RET, MET, CDK4, and KIT that have been found causative in cancer predisposition syndromes multiple endocrine neoplasia type 2 (MEN2) (Donis-Keller et al. 1993, Mulligan et al. 1993), hereditary papillary renal carcinoma (HPRC) (Zuo et al. 1996), familial malignant melanoma (FMM) (Schmidt et al. 1997), and familial gastrointestinal stromal tumour (GIST) syndrome (Nishida et al. 1998), respectively.

1.3 Epigenetic factors

Tumorigenic processes can also advance through epigenetic changes that affect gene expression levels without directly altering the DNA sequence. Epigenetic events such as altered methylation patterns, post-translational modification of histones, or arrangement of nucleosomes can lead to inactivation of tumour suppressor genes,
activation of oncogenes, or altered imprinting patterns (Lund and van Lohuizen 2004).

A prime example of an epigenetic event affecting tumour formation is the hypermethylation of MLH1 promoter frequently observed in sporadic CRCs with microsatellite instability (MSI) (Kane et al. 1997, Cunningham et al. 1998, Herman et al. 1998, Veigl et al. 1998). Hypermethylation seems to be the main mechanism of MLH1 inactivation in sporadic MSI CRCs (Kane et al. 1997, Cunningham et al. 1998, Herman et al. 1998, Veigl et al. 1998). MLH1 inactivation abrogates normal MMR functions and results in a mutator phenotype due to the accumulation of microsatellite mutations. Other genes, for example MGMT and CDKN2A are also frequently hypermethylated already in adenomas of the colon (Petko et al. 2005).

Regulation of gene expression via loss of imprinting (LOI) has also been implicated in CRC (Kinouchi et al. 1996). Retained biallelic expression of IGF2 has been linked to uncontrolled cell growth (Ogawa et al. 1993) and increased CRC risk (Cui et al. 2003).

2. Cancer susceptibility

Although a majority of cancer cases occur sporadically, hereditary counterparts of many cancer types have been identified (Vogelstein and Kinzler 2004). The risk of developing cancer is higher in individuals with an inherited predisposition. The total risk is dependent on the mode of inheritance, penetrance, contributing genetic factors (modifier genes), as well as on dietary, lifestyle, and environmental factors. Most of the cancer susceptibility syndromes have an autosomal dominant mode of inheritance but some rare examples of autosomal recessive syndromes exist, such as the MYH-associated colorectal polyposis (Al-Tassan et al. 2002). Inheriting a high-penetrance susceptibility allele increases cancer risk per se. Cancer families with mutations in high-penetrance genes are often characterized by: 1. younger age at onset than patients with sporadic disease, 2. bilateral disease or multiple disease sites in one organ, 3. multiple primary malignancies, 4. occurrence of cancer in the less usually affected sex, 5. cancer clustering in the family, and 6. cancer patients displaying other rare conditions (Marsh and Zori 2002). In comparison, the risk conferred by a low-penetrance susceptibility allele is influenced more by other genetic alterations and environmental factors. Therefore cancer usually manifests later and the degree of familial clustering is lower than for high-penetrance mutations (Kemp et al. 2004).

A single gene can harbour both high- and low-penetrance susceptibility alleles, as exemplified by the different sequence alterations in APC associated with diverse CRC risks and phenotypic manifestations. Different APC alterations can predispose either to: 1. a numerous amount of adenomatous polyps in familial adenomatous polyposis (FAP) patients with a near 100% risk of developing CRC if untreated (Kinzler and Vogelstein 1996), 2. a less severe polyposis phenotype and CRC risk in attenuated familial adenomatous polyposis (AFAP) patients (Kinzler and Vogelstein 1996), or 3. a modest, 1.5-2 fold increased CRC risk without an associated polyposis phenotype for example in Ashkenazi Jewish populations (Locker and Lynch 2004).
First-degree relatives of cancer patients with many common cancer types are at a 2-3 fold increased risk of developing cancer at the same site (Peto and Houlston 2001). Much of this familial aggregation is due to inherited susceptibility (Lichtenstein et al. 2000). Lichtenstein et al. (2000) estimated the heritability of cancer at 28 anatomical sites by studying almost 45000 twin pairs. The contribution of heritable factors was highest in prostate, colorectal, and breast cancer. An estimated 42% of the total prostate cancer risk was attributed to heritable factors, and the estimates for colorectal and breast cancer were 35% and 27%, respectively (Lichtenstein et al. 2000).

3. Familial colorectal cancer

Colorectal cancer (CRC) is the third most common cancer type and the fourth most common cause of cancer deaths worldwide (Parkin et al. 2005). CRC is more prevalent in the Western countries possibly due to environmental and dietary factors (Potter 1999). Hereditary factors have an important role in CRC with an estimated 35% of the total risk attributed to inherited susceptibility (Lichtenstein et al. 2000).

To date, a number of CRC susceptibility syndromes have been characterized that can be divided into two main groups based on the presence or absence of polyposis. Polyposis syndromes can be further divided into adenomatous and hamartomatous polyposis based on the disease phenotype. MYH-associated polyposis, familial adenomatous polyposis (FAP), and its variants attenuated familial adenomatous polyposis (AFAP), Gardner syndrome, and Turcot syndrome (TS) are characterized by adenomatous polyposis. Hamartomatous polyposis syndromes juvenile polyposis (JP), Peutz-Jeghers syndrome (PJS), Cowden syndrome (CS), and its variant Bannayan-Riley-Ruvalcaba syndrome have lower cancer risk than in adenomatous polyposis syndromes. The non-polyposis syndromes hereditary non-polyposis colorectal cancer (HNPCC) and its variants Muir-Torre syndrome (MTS) and Turcot syndrome are typically devoid of florid polyposis.

The most common and best characterized hereditary CRC syndromes FAP and HNPCC will be described in more detail in the following sections.

3.1 Familial adenomatous polyposis, FAP

FAP is characterized by the formation of hundreds, even thousands of adenomatous polyps and microadenomas in the large intestine already in adolescence or early adulthood. If untreated, colorectal cancer formation is inevitable by the early forties due to florid polyposis (Fearnhead et al. 2001). FAP is caused by mutations in APC (Groden et al. 1991, Nishisho et al. 1991). APC acts as a negative regulator of β-catenin in the WNT signalling pathway (Munemitsu et al. 1995). β-catenin in turn interacts with TCF and LEF transcription factors that control the expression levels of many genes, including oncogene c-Myc (He et al. 1998). Most APC mutations result in C-terminally truncated protein products with abrogated β-catenin binding (Kinzler and Vogelstein 1996). This leads to constitutive upregulation of the WNT signalling target genes (Korinek et al. 1997). Furthermore, truncated APC cannot promote the assembly of microtubules (Munemitsu et al. 1994, Smith et al. 1994) and may thus contribute to chromosomal instability (Fodde et al. 2001, Kaplan et al. 2001).
FAP patients display highly variable clinicopathological characteristics. For example, the number of polyps, age at cancer onset, and extracolonic manifestations such as the occurrence of desmoid tumours can vary between but also within families. This can be partly explained by the mutation spectrum of \textit{APC}. For example, mutations leading to severe polyposis phenotype and early age at onset occur mostly in the central region of the gene (Nagase et al. 1992, Caspari et al. 1994, Gayther et al. 1994). In comparison, attenuated FAP is associated with mutations in the first or last third of \textit{APC} (Spiro et al. 1993, Friedl et al. 1996, van der Luijt et al. 1996). The phenotype of FAP patients with the same \textit{APC} mutation may also vary within and between families. This could be accounted for by inaccurate diagnosis, variability in the micro- or macroenvironmental conditions, or modifier genes (Houlston et al. 2001, Crabtree et al. 2002).

3.1.1 FAP modifier genes

Crabtree et al. (2002) compared phenotypes between close and more distant relatives with FAP and found evidence to support the existence of genetic modifiers of the disease. The modifier hypothesis is further supported by studies on \textit{Min} (multiple intestinal neoplasia) mouse model of FAP. \textit{Mom1} and \textit{Mom2}, the modifier loci of \textit{Min}, have been localized by linkage analysis to mouse chromosomes 4 and 18, respectively (Dietrich et al. 1993, Silverman et al. 2002). No evident mouse modifiers in \textit{Mom2} locus have been identified to date (Silverman et al. 2003) but in \textit{Mom1} locus, the variants of \textit{Pla2g2a} (\textit{Pla2s}) and \textit{Myh} have been seen to modify polyp number (MacPhee et al. 1995, Sieber et al. 2004). However, the human homologs \textit{PLA2G2A} or \textit{MYH} do not seem to modify the FAP phenotype (Dobbie et al. 1996, Tomlinson et al. 1996, Plasilova et al. 2004, Kairupan et al. 2005). In humans, variants of N-acetyltransferases \textit{NAT1} and \textit{NAT2} have been suggested to modify the FAP phenotype (Crabtree et al. 2004).

3.2 Hereditary non-polyposis colorectal cancer, HNPPC

HNPPC patients are typically devoid of florid polyposis and adenoma development seems to be as frequent as in the general population. The aggressive adenoma theory of Jass (1995) suggests that the adenomas progress into carcinomas more rapidly and/or more often than in the general population. The culprit behind this accelerated tumour formation in HNPPC patients is defective mismatch repair (MMR) system unable to correct the small insertion/deletion mutations frequently generated during the replication of repeat sequences. Because the mutations cannot be corrected, malignant transformation is promoted because of accumulating microsatellite mutations (microsatellite instability, MSI) in genes involved in cell growth, differentiation, and apoptosis.

and $MSH2$ (Peltomäki and Vasen 2004). Although most mutations (81%) are unique and specific to each family (Peltomäki and Vasen 2004), there are some founder mutations in, for example, the Ashkenazi Jews (Foulkes et al. 2002), Finns (Nyström-Lahti et al. 1995), and Newfoundlanders (Froggatt et al. 1999). Altogether 13% of the mutation positive cases seem to be associated with three of the most common $MLH1$ and $MSH2$ mutations (Peltomäki and Vasen 2004).

HNPCC is a somewhat misleading designation due to the occurrence of cancer in a number of other tissue types as well. Hence, the name Lynch syndrome has been used in parallel. The syndrome is associated with an increased cancer risk of, for example endometrium, stomach, ovary, ureter/renal pelvis, small bowel, and hepatobiliary tract (Watson and Lynch 1993, Aarnio et al. 1995). The risk of developing CRC or endometrial cancer seems to be highest. Cumulative CRC risk values ranging from 54% to 100% have been obtained, partly depending on the mutated gene in question, and endometrial cancer risk estimations have varied between 24% and 62% (Quehenberger et al. 2005). Quehenberger et al. (2005) suggest, however, that the previously presented CRC risks were overestimations due to the inclusion of mostly high risk families in the analyses.

Phenotypic manifestations in a given HNPCC family seem to be dependent on the mutant MMR gene in question. $MSH2$ mutation carriers might be at higher risk of developing cancer at any site than $MLH1$ mutation carriers (Vasen et al. 1996, Vasen et al. 2001). CRC risk associated with $MSH6$ seems to be lower than for $MLH1$ or $MSH2$, but the risk of endometrial cancer could be higher (Wijnen et al. 1999, Hendriks et al. 2004). Mutations in $PMS2$ have been previously identified mostly in families with features of Turcot syndrome (Peltomäki and Vasen 2004) but were recently identified also in a bone fide HNPCC family (Worthley et al. 2005). Similar to FAP patients, HNPCC patients with identical mutations can display different disease phenotypes indicating the importance of environmental factors and/or modifier genes.

The role of MMR genes other than $MLH1$, $MSH2$, $PMS2$, and $MSH6$ has been considered since a fraction of the HNPCC cases do not seem to associate with these genes (Liu et al. 1996). The implications of $PMS1$, $MLH3$, and $MSH3$ mutations are however not clear. It has also been postulated that using expression-based mutation detection methods, the known predisposition genes might be found to be causative in a fraction of the previously seemingly mutation negative cases (Renkonen et al. 2003).

3.2.1 Amsterdam and Bethesda criteria

Amsterdam criteria I were proposed in 1991 to facilitate the identification of HNPCC families (Vasen et al. 1991). In 1999, the Amsterdam criteria I were supplemented to contain the extracolonic features associated with HNPCC (Vasen et al. 1999). The Amsterdam criteria II are as follows:

At least three family members should be affected with an HNPCC-associated cancer (CRC, cancer of the endometrium, ureter, renal pelvis, or small bowel) and in addition, the following criteria should be met:
- one patient should be a first-degree relative of the other two patients
- at least two successive generations should be affected
- at least one of the associated cancers should be diagnosed before age 50
- FAP should be excluded
- pathological examinations should verify the tumours

The Bethesda guidelines were introduced in 1997 to help specifying the tumours that should be tested for MSI, a typical characteristic of HNPCC. The Bethesda criteria are based on age at diagnosis, pathological characteristics, family history, and the presence of extracolonic tumours (Rodriguez-Bigas et al. 1997). In 1998, a Bethesda panel consisting of five microsatellite markers BAT25, BAT26, D5S346, D2S123, and D17S250 was introduced to unify MSI-testing (Boland et al. 1998). Interpretation of the results obtained with the Bethesda panel was proposed as following:

- if two or more markers display instability, high level MSI (MSI-H) is scored
- if instability is detected with one marker, low level MSI (MSI-L) is scored
- if all markers are stable, microsatellite stability (MSS) is scored

In case of MSI-L, additional markers should be analyzed. MSI-L could be scored if less than 30% of the markers show MSI (Boland et al. 1998).

The revised Bethesda guidelines introduced in 2004 suggest that more mononucleotide repeats could be analyzed to increase the detection sensitivity of MSI-H tumours (Umar et al. 2004). Furthermore, MSI-L specific markers, such as MYCL, could be used. However, the division of CRCs into three distinct groups based on the degree of microsatellite instability has been a matter of debate (Umar et al. 2004). This issue will be discussed in the following chapter.

4. Genetic instability in colorectal cancer

CRC arises from a stem cell in the colonic crypt and the earliest identifiable lesion is an aberrant crypt focus (ACF). The progression into carcinoma proceeds via two main pathways, chromosomal instability (CIN) pathway or the microsatellite instability (MSI) pathway (Vogelstein et al. 1988, Fearon and Vogelstein 1990, Aaltonen et al. 1993, Ionov et al. 1993, Thibodeau et al. 1993). Typically around 85% of all CRC cases display CIN with gross chromosomal aberrations whereas the remaining 15% display MSI with more subtle aberrations in microsatellites. FAP is a typical example of a CRC syndrome displaying CIN whereas MSI is a typical feature of HNPCC.

A mild MSI phenotype (MSI-L) characterizes a subset of CRC cases and they have been suggested to constitute a distinct entity. Attempts to differentiate MSI-L from MSS tumours according to their clinicopathological or molecular features have produced conflicting results (Thibodeau et al. 1998, Jass et al. 1999, Halford et al. 2002, Laiho et al. 2002). It has also been suggested that if enough microsatellite markers are analyzed, almost all CRCs display MSI-L (Halford et al. 2002, Laiho et al. 2002). However, a subset of MSI-L CRCs seems to display distinct morphological features including a typical saw-toothed or serrated epithelial structure. This epithelial morphology characterizes also a small subset of MSS tumours suggesting the heterogeneity of the so-called serrated pathway (Sawyer et al. 2002). The pathway
depicts the progression of hyperplastic polyps to serrated adenomas and further to serrated CRCs (Iino et al. 1999, Jass 1999, Hawkins and Ward 2001). High incidence of \textit{BRAF} mutations in serrated polyps proposes the early involvement of \textit{BRAF} in the serrated neoplasia pathway (Chan \textit{et al.} 2003). An expression analysis on serrated CRCs suggests that they form not only morphologically but also biologically distinct class of CRCs (Laiho \textit{et al}, unpublished data).

4.1 Features of colorectal tumours arising through the CIN and MSI pathways

MSS or CIN CRCs are typically aneuploid, left-sided, behave aggressively, and harbour characteristic mutations in, for example \textit{KRAS}, \textit{APC}, and \textit{TP53}. In contrast, MSI CRCs are predominantly diploid, right-sided, poorly differentiated, often characterized by a strong lymphocyte infiltration, and harbour distinct mutations in, for example \textit{TGF\textbeta RII} and \textit{BAX} (Smyrk and Lynch 1999). Most of the published studies suggest a better prognosis for patients with MSI CRC (Pawlik \textit{et al.} 2004).

The clinicopathological and molecular differences between proximal and distal CRCs might be partly explained by the different embryological origin and blood vasculature of the colonic components. Differences in the colonic microenvironment might cause a different cellular response to different genetic changes and environmental factors (Bufill 1990, Bardelli \textit{et al.} 2001, Glebov \textit{et al.} 2003).

4.2 Adenoma-carcinoma sequence

The clonal progression of CIN CRCs is depicted by the adenoma-carcinoma sequence (Figure 1) (Vogelstein \textit{et al.} 1988, Fearon and Vogelstein 1990). Multiple alterations in both oncogenes and tumour suppressor genes are required for malignant transformation. Although the alterations often seem to follow a distinct sequence, the number of changes seems to be the critical aspect rather than the actual order of the molecular alterations (Fearon and Vogelstein 1990).

The primary alterations in the adenoma-carcinoma sequence affect \textit{APC}. \textit{APC} mutations or 5q LOH targeting \textit{APC} have been detected already in a large proportion of early adenomas, but equally frequently also in carcinomas (Vogelstein \textit{et al.} 1988, Powell \textit{et al.} 1992, Leslie \textit{et al.} 2002). Mutations in oncogene \textit{KRAS} are found equally frequently in large adenomas and carcinomas, but less frequently in small adenomas suggesting the role of \textit{KRAS} in promoting growth rather than initiating tumorigenesis (Vogelstein \textit{et al.} 1988, Leslie \textit{et al.} 2002). KRas is involved in controlling cellular proliferation and differentiation pathways. In addition to 5q deletions, 18q LOH is a frequent event in CRC tumorigenesis. Approximately 70\% of carcinomas and a smaller proportion of early adenomas harbour 18q deletions (Vogelstein \textit{et al.} 1988, Boland \textit{et al.} 1995, Leslie \textit{et al.} 2002). The most likely target genes for 18q deletions are \textit{SMAD2} and \textit{SMAD4} or the previously proposed \textit{DCC} (Leslie \textit{et al.} 2002). Both SMADs are involved in the TGF-\textbeta signalling pathway regulating for example cell growth, differentiation, and apoptosis (Heldin \textit{et al.} 1997). Loss of chromosome 17p that targets \textit{TP53} is also frequently observed in CRCs (Vogelstein \textit{et al.} 1988, Leslie \textit{et al.} 2002). \textit{TP53}}
alterations are more frequent in carcinomas than in adenomas indicating the possible association of defective p53 in adenoma-carcinoma transition. p53 maintains genomic stability by controlling cell cycle progression and apoptosis.

Progression into metastatic CRC requires additional molecular changes in order for the tumour to invade surrounding tissues, be released into the circulation, and colonize new sites such as the liver. The exact molecular events controlling CRC metastasis are not fully known. The involvement of, for example, PRL-3 and multiple factors in the WNT/β-catenin pathway has been suggested (Brabletz et al. 2001, Rivat et al. 2003, Zeng et al. 2003, Pai et al. 2004, Dhawan et al. 2005). Furthermore, recent protein expression comparisons between primary CRCs and hepatic metastases identified 9 differentially expressed proteins, such as proapolipoprotein and beta-globin, with potential involvement in carcinogenesis and metastatic processes (Yu et al. 2004).

![Figure 1](image-url) The sequential progression model of CRC tumorigenesis. Modified from Fearon and Vogelstein (1990) and Kinzler and Vogelstein (1996).

The exact mechanisms initiating CIN are not well known and over 100 candidate genes have been proposed in yeast (Lengauer et al. 1998, Jallepalli and Lengauer 2001). Many of the candidate genes are involved in cell cycle checkpoints and some of these genes, for example RAD17 and AURKA have been linked to tumorigenic processes based on the observed overexpression in CRCs (Bischoff et al. 1998, Bao et al. 1999). Furthermore, BUB1 and BUBR1 mutations have been observed in CRCs (Cahill et al. 1998). In addition to the checkpoint genes, telomere integrity is known to suppress CIN.

4.3 Microsatellite instability pathway

An alternative pathway of tumorigenesis is the microsatellite instability (MSI) pathway caused by defective DNA mismatch repair system (MMR) (Figure 2). Cells with inactive MMR genes, mostly MLH1 and MSH2, have approximately 100-fold higher mutation frequencies (Ionov et al. 1993, Strand et al. 1993, Bhattacharyya et al. 1994, Shibata et al. 1994). The mutations following MMR inactivation can affect genes important in for example cell cycle regulation and DNA repair. These events are required for malignant transformation and occur before visible tumour formation (Tsao et al. 2000, Shibata 2001). The number of microsatellite mutations can reflect the pattern of tumour progression and serve as molecular tumour clocks (Shibata et al. 1996).
There are several potential mutational targets in MMR deficient cells since microsatellites are abundantly found in coding and non-coding regions of the human genome (Toth et al. 2000). Microsatellite mutations have been observed in a number of putative MSI target genes and the tumorigenic implications of these mutations have been presented in some cases, such as TGFβRII and BAX (Wang et al. 1995, Ionov et al. 2000). Current knowledge on MSI target genes will be discussed in more detail in section 4.3.2.

**Figure 2.** Microsatellite-unstable colorectal cancer progression.

4.3.1 DNA mismatch repair, MMR

MMR system maintains genomic stability mainly by correcting single base mismatches and small insertion/deletion loops (IDLs) that escape the nucleotide selection and proofreading activities of DNA polymerases during DNA replication. MMR proteins are also involved in: 1. inhibiting recombination of divergent DNA sequences, 2. correcting errors caused by mutagens such as alkylating agents, 3. participating in double-strand break repair as well as transcription coupled repair together with NER proteins, and 4. regulating DNA damage signalling and apoptosis (Harfe and Jinks-Robertson 2000).

MMR mechanisms are highly conserved in eukaryotes and prokaryotes. The *E.coli* MMR system is the best characterized and multiple MutS, MutL, and MutH proteins are known to function in mismatch recognition, mediating recognition and repair processes, and targeting repair to the newly synthesized DNA strand, respectively. All eukaryotic organisms studied to date have been shown to possess multiple MutS homologs (MSH proteins) and MutL homologs (MLH proteins) but no convincing MutH homologs have been identified (Harfe and Jinks-Robertson 2000).

In human error repair systems, mismatches are recognized by the MSH2/MSH6 complex (MutSα) and IDL recognition is mediated through MutSα and the MSH2/MSH3 (MutSβ) complex. Subsequently, MutS complexes interact with MutL complexes MLH1/PMS2 (MutLα) and occasionally with MLH1/MLH3 or MLH1/PMS1 (MutLβ). The MutL complexes mediate the error recognition to the downstream events involved in error repair (Kolodner and Marsischky 1999, Marra and Schar 1999, Harfe and Jinks-Robertson 2000) (Figure 3). The mechanism used to recognize the newly synthesized strand in human cells is not known. In *E.coli* the synthesized strand is transiently undermethylated but in humans the methylation patterns are irregular. It has been postulated that replication of the lagging strand as Okazaki fragments and formation of nicks during the process could aid in the discrimination of the DNA strands. The leading strand could be identified by the growing 3’ end. The role of PCNA in strand recognition has also been suggested due to the observed interaction with MMR proteins.
Figure 3. Principal proteins involved in DNA mismatch repair in humans. Interactions between MutS (MSH2/MSH6/MSH3) and MutL (MLH1/PMS2) homologs mediate the recognition of base mismatches and insertion/deletion loops. Modified from Marra and Schär (1999).

The excision of the newly synthesized DNA fragment containing the error involves DNA exonuclease(s) and at least EXO1 has been linked to human MMR system (Tishkoff et al. 1997). Resynthesis of the excised strand is mediated by DNA polymerases and replication factors. DNA ligases subsequently complete the MMR processes.

The functional redundancy of MMR proteins in error recognition has an impact on the MSI status of tumours with mutations in different MMR genes. MLH1 or MSH2 deficient cells display high degree of MSI because of the involvement of these proteins in all MutL and MutS complexes, respectively. However, the variable degree of MSI in MSH6 deficient tumours is explained by the ability of MSH2/MSH3 complex to compensate for the IDL correction.

Interestingly, homozygous germline mutations of MMR genes MLH1 and MSH2 have been linked to neurofibromatosis type 1 and early onset haematological malignancy (Ricciardone et al. 1999, Wang et al. 1999, Whiteside et al. 2002). Defective MMR has been shown to render NF1 susceptible to somatic frameshift mutations (Wang et al. 2003). Furthermore, a woman heterozygous for two different MLH1 missense mutations developed breast cancer at the age of 35. The mild phenotype observed in the patient might be due to a residual MLH1 activity from the other allele (Hackman et al. 1997).

4.3.2 MSI target genes

Cells with defective MMR system accumulate microsatellite mutations in both noncoding and coding regions of the genome. If genes involved in controlling cell growth or maintaining genomic integrity are affected, tumour progression may be promoted. Generally these consequential mutations occur in coding microsatellites, leading to the formation of altered protein products or their degradation via nonsense-mediated decay (NMD). In addition to coding region mutations, mutations occurring
in regulatory non-coding regions may also have an impact on the gene functions. Usually, however, non-coding microsatellite mutations are thought to be inconsequential background events.

High mutation frequency has been considered as a useful criterion for defining true MSI target genes (Boland et al. 1998). Other criteria suggested by Boland et al. (1998) included functional evidence, biallelic inactivation, role in a growth suppressor pathway, and inactivation of the same growth suppressor pathway in MSS tumours. Some of these criteria have been questioned since. Monoallelic mutations might suffice in case of haploinsufficiency or dominant negative effect (Liu et al. 2000, Yamamoto et al. 2000). The affected genes might also have roles other than tumour suppressor functions, as exemplified by somatic microsatellite mutations observed in MMR genes *MSH3* and *MSH6*. Of note, somatic inactivation of *MSH3* and *MSH6* has been shown to further increase the instability phenotype (Akiyama et al. 1997b, Baranovskaya et al. 2001, Duval et al. 2001). Furthermore, the pathways involved in MSS and MSI tumours seem to differ. Hence, the most widely accepted criteria for defining true MSI target genes have been high mutation frequency and functional evidence.

Duval and Hamelin (2002) have suggested that genes mutated in MSI tumours could be divided into four groups based on the observed mutation frequency and the roles of the protein products. Survivor genes (1.) with essential cellular roles would be rarely mutated due to negative selection pressure. Hibernator genes (2.) with unimportant functions would accumulate mutations at the background level. The effect of co-operator gene (3.) mutations would be emphasized when occurring in parallel with mutations in other co-operator genes. Their mutation frequency would therefore vary depending on the context in which they occur. The highest mutation frequencies would be observed in transformator genes (4.) because the mutations confer a growth advantage (Duval and Hamelin 2002).

The mutation frequencies of true MSI target genes can thus vary extensively depending on their role in the affected cells. Furthermore, the mutation frequency of a given microsatellite seems to be dependent on the repeat type and length, as well as yet unknown influence from the surrounding sequence (Zhang et al. 2001, Suzuki et al. 2002, Vilikki et al. 2002). It is important to study the general background mutation frequency of MSI tumours to obtain information on the possible sequence elements affecting replication fidelity. In addition, functional evidence has to be provided to prove the importance of a given gene in MSI tumorigenesis (Perucho 2003).

To date, MSI tumours have been analyzed for microsatellite mutations in a large number of genes, but only a handful of genes have been shown to contribute to the MSI pathway of tumour progression. The first gene identified with somatic microsatellite mutations was *TGFβRII*. Subsequently, around 90% of MSI CRCs have been shown to harbour *TGFβRII* mutations, often in a biallelic form (Parsons et al. 1995). Functional studies have indicated a loss of *TGFβRII* tumour suppressor function as a result of these repeat mutations (Wang et al. 1995). Furthermore, *TGFβRII* mutations seem to be an early event in MSI tumorigenesis (Grady et al. 1998, Duval et al. 2001). In addition to *TGFβRII*, mutations in for example proapoptotic factor *BAX* and growth factor receptor *IGFRII* have been shown to contribute to MSI tumorigenesis (Souza et al. 1999, Ionov et al. 2000).
5. Prospects of CRC studies

The thus far identified hereditary polyposis and non-polyposis syndromes together account for approximately 5% of all CRC cases (Burt and Neklason 2005). Novel CRC susceptibility genes are still to be identified and evidence for putative loci on 9q and 15q has been obtained (Tomlinson et al. 1999, Jaeger et al. 2003, Wiesner et al. 2003). The underlying high penetrant mutations are most likely rare in the general population and are therefore causative in a minority of all CRC cases. In comparison, the more frequent low-penetrance susceptibility alleles might contribute to a larger proportion of all CRC cases.

Low-penetrance genes confer a modestly increased cancer risk and mild clustering of disease in the family. Therefore, the identification of causative genetic alterations needs to be pursued by methods other than linkage, a successfully utilized method in the identification of high-penetrance susceptibility genes (Houlston and Peto 2004). Allelic association analyses of low-penetrance genes can be performed with much smaller sample size than linkage analyses (Risch and Merikangas 1996, Camp 1997) and the number of cases needed for association analysis would be even lower if affected relatives were included in the analyses (Houlston and Peto 2003). In fact, association analysis has been the most common method in identifying low-penetrance susceptibility alleles (Houlston and Tomlinson 2001).

5.1 Candidate low-penetrance CRC susceptibility alleles

To date, multiple association analyses have been performed to estimate the CRC risk conferred by sequence variants in different genes. With increasing numbers of association analyses, some pooled analyses have also been undertaken to increase the number of samples per polymorphism analyzed (Houlston and Tomlinson 2001, de Jong et al. 2002, Chen et al. 2005).

One meta-analysis suggested APC I1307K as a candidate low-penetrance susceptibility allele (Houlston and Tomlinson 2001). The APC I1307K change is more frequent in CRC patients than in healthy controls in Ashkenazi Jewish populations. The amino acid change creates a hypermutable poly-A tract that is easily affected by a deletion or amplification event (Laken et al. 1997). In addition to APC I1307K, an association between HRAS-VNTR as well as MTHFR variants and CRC risk has been suggested (Houlston and Tomlinson 2001, de Jong et al. 2002). HRAS oncogene is involved in the mitogenic signal transduction and differentiation. Variable number of tandem repeats (VNTR) located ~1kb downstream of HRAS has been shown to interact with transcriptional regulatory elements and control the expression levels of nearby genes. MTHFR in turn catalyzes the conversion of 5,10-methylenetetrahydrofolate to 5-methyltetrahydrofolate which is the primary circulatory form of folate. The availability of folate and other methyl group donors has been suggested to affect methylation (Kim et al. 1996).

The role of GSTT1 genotype and NAT2 phenotype has also been suggested in two meta-analyses (de Jong et al. 2002, Chen et al. 2005). GSTT1 and NAT2 are carcinogen metabolism enzymes mediating defenses against a number of chemical carcinogens and environmental toxins (Upton et al. 2001, Hayes et al. 2005).
5.2 Identification of novel low-penetrance susceptibility alleles

A number of association studies have been performed to identify low-penetrance CRC susceptibility alleles. The relative risk attributed to a given polymorphism varies between the analyses and the meta-analyses discussed above. For example, the role of MTHFR alleles in CRC was supported by Houlston and Tomlinson (2001) and de Jong et al. (2002) but questioned by Chen et al. (2005). This may be due to different composition of studies included in the meta-analysis. Furthermore, there may be differences in the variables, for example ethnicity, gender, and tumour localization, that were taken into account when assessing the relative risks. When assessing the importance of a given polymorphism based on current literature, it has to be remembered that publication bias might have hindered the publication of negative results.

The limitation of association analysis is that in the absence of a good candidate polymorphism, the number of associations tested can be prohibitively large and the problem of multiple testing also rises. The number of markers needed for association analysis with sufficient power is dependent on the extent of linkage disequilibrium (LD) between a marker and a causal locus. Typically LD extends ~50-60kb in the human genome but even further in populations with homogeneous gene pool (Reich et al. 2001). The extent of LD can also be higher in young admixed populations that have been formed by gene flow from two or more genetically distinct founding populations (Stephens et al. 1994). Mixing of genetic material creates LD between linked but also unlinked loci that have different allele frequencies in the founding populations. Therefore, methods to discriminate between spurious and true associations need to be utilized (Pfaff et al. 2001).

Identification of low-penetrance cancer susceptibility genes can benefit from advances in gene-mapping techniques in mice. Crosses between susceptible and resistant strains have produced information on several mouse tumour susceptibility loci (Demant 2003). Furthermore, studying the interactions between multiple polymorphisms is much more straightforward in mice (Mao and Balmain 2003). However, information obtained from mouse studies might not be applicable in humans due to inter-species differences. This is exemplified by the discrepancies between mouse and human studies on modifying factors of FAP, as discussed in section 3.1.1.

For the reasons stated above, the identification of low-penetrance susceptibility alleles is understandably a laborious effort. Since these alleles can occur frequently in the general population, they can have a vast impact on general public health. It is therefore important to study the genotype-phenotype correlations and utilize novel technologies, for example dense SNP maps, SNP arrays, and array-CGH in the identification of the low-penetrance susceptibility alleles. Furthermore, prior information on potentially interesting genomic positions or gene functions would decrease the number of associations tested. A candidate genomic region may be assigned by utilizing chromosomal clues of for example commonly deleted or amplified genomic regions. Emerging novel technologies combined with novel information from cellular pathways can facilitate the identification of low-penetrance susceptibility alleles.
AIMS OF THE STUDY

Novel genes and loci associated with microsatellite stable colorectal cancer predisposition

I Identification of novel susceptibility loci in familial colorectal cancer.

II Role of AURKA in colorectal cancer susceptibility: a candidate gene selected based on study I.

Causes and characteristics of colorectal cancers displaying microsatellite instability

III Evaluation of the putative role of MLH3 in microsatellite-unstable colorectal cancer.

IV Assessment of SEMG1 as a candidate microsatellite instability target gene and evaluation of the general background mutation frequency of microsatellite-unstable colorectal cancers.
MATERIALS AND METHODS

1. Colorectal cancer patients and DNA samples

The colorectal carcinoma and corresponding normal tissue samples utilized in studies I-IV were chosen from sample series collected since 1994 (Aaltonen et al. 1998, Salovaara et al. 2000, Laiho et al. unpublished data). The series contain samples from altogether 1579 patients (at the time of the thesis’ last project). Information on the family background was obtained from official population registries and cancer occurrence in the family from the Finnish Cancer Registries. A majority (86%) of the patients have no first-degree relatives affected with CRC. Familial cases were defined as having at least one first-degree relative affected with CRC. Most of the families display a mild CRC clustering. The samples and patient information were obtained with approval from the Ethics committees of the Department of Medical Genetics, University of Helsinki, Helsinki University Central Hospital, and Hospital District of Helsinki and Uusimaa.

The normal tissue DNA was extracted from blood or normal colonic epithelium distant from the tumour margins. The fresh frozen tumour tissue samples were histologically evaluated by a pathologist before DNA extraction. Each tumour sample undergoes an MSI analysis and of the currently established MSI cases, 22% (35/162) harbour germline MLH1, MSH2, or MSH6 mutations (Aaltonen et al. 1998, Salovaara et al. 2000, Loukola et al. 2001).

1.1 Common regions of allelic imbalance (I)

Normal and tumour tissue samples were analyzed with microsatellite markers to detect common genomic regions of loss of heterozygosity/amplification (allelic imbalance, AI). Samples from altogether 99 familial and 186 sporadic Finnish CRC patients were utilized in study I. Polyposis cases (FAP, PJS, and JP) as well as HNPCC and other cases with MSI were excluded from the study. All tumour samples displayed at least 50%, and typically ~75%, of carcinoma tissue.

Additional normal tissue and CRC samples obtained from collaborators in Denmark and UK were analyzed for AI with microsatellite marker D20S178. Samples from altogether 18 familial and 24 sporadic cases from Denmark, and 67 familial and 96 sporadic cases from the UK were included in study I.

1.2 Genetic analyses of AURKA (II)

Normal and tumour tissue DNA samples from 125 familial and 110 sporadic Finnish MSS CRC cases were analyzed. Polyposis cases (FAP, PJS, JP, and MYH-associated polyposis) as well as HNPCC and other cases displaying MSI were excluded from the study. A majority (68%) of the cases had been analyzed for AI in study I. A majority of the tumour samples (91%) displayed at least 60% carcinoma tissue.
In addition, samples from 94 anonymous cancer-free blood donors provided by the Finnish Red Cross Blood Transfusion Centre were utilized.

1.3 *MLH3* mutation analysis (III)

Normal tissue DNA samples from 30 Finnish CRC patients displaying MSI were included in study III. The samples were chosen based on the molecular features suggested to be causally associated with *MLH3* defects (Wu et al. 2001). Namely, samples from 17 familial and 6 sporadic CRC patients were included, all of which were stable when analyzed with mononucleotide repeat marker BAT26. In addition, samples from all 7 available mutation-negative HNPCC patients were analyzed. The mutation negative HNPCC samples were provided by Professor Päivi Peltonäki. A tumour tissue sample from a patient with germline *MLH3* Met809Val mutation was analyzed. The tumour tissue sample displayed 70% carcinoma tissue.

In addition, samples from 713 anonymous cancer-free blood donors provided by the Finnish Red Cross Blood Transfusion Centre were utilized.

1.4 Genetic analyses of *SEMG1* and intergenic T9 repeats (IV)

Tumour tissue samples from 146 MSI CRC patients were utilized in study IV. The respective normal tissue samples available from 144 patients were also analyzed. In addition, MSS CRC samples from 238 patients and normal tissue samples from 50 cancer-free blood donors from the Finnish Red Cross Blood Transfusion Centre were utilized.

Most of the tumour samples used in study IV (92%) contained ≥ 50% carcinoma cells. Altogether 20% of the MSI CRC patients included in study IV harbour germline *MLH1* or *MSH2* mutations (Aaltonen et al. 1998, Salovaara et al. 2000, Loukola et al. 2001).

2. Cancer cell lines (IV)

In study IV, we utilized 31 cancer cell lines originating from colon (n=28), prostate (n=2), and endometrium (n=1). Of the CRC cell lines, 10 displayed MSI and 18 were MSS. Both prostate cancer cell lines as well as the endometrial cancer cell line displayed MSI.

3. Analysis methods

3.1 Common regions of allelic imbalance (I)
3.1.1 Microsatellite marker analysis

The AI frequency of familial and sporadic tumour sets was determined by using microsatellite marker analysis. Normal and tumour tissue samples from 29 familial and 75 sporadic CRC patients were analyzed with 372 microsatellite markers (ABI PRISM Linkage Mapping Set MD-10, P/N 450067, 10 cM density, Applied Biosystems (AB) Division, Foster City, CA). The DNA samples were amplified using fluorescent labelled primers and run on an ABI PRISM 377 DNA Sequencer (AB) according to manufacturer’s instructions. The fragments were analyzed using GeneScan 3.1 and Genotyper 2.5 programs (AB). The samples displaying AI were determined using the previously published mathematical model (Canzian et al. 1996). In effect, AI was scored if the other allele emitted at least 40% less fluorescence. From this primary dataset of 372 markers, those markers showing the most promising association of AI and familial CRC were further analyzed using additional normal and tumour tissue samples from 70 familial and 111 sporadic CRC patients.

The samples obtained from collaborators in Denmark and UK were analyzed with one specific microsatellite marker, D20S178, to determine the AI frequencies in the familial and sporadic groups.

3.1.2 Comparative genomic hybridization

Data from a previously performed, unpublished CGH analysis was utilized to obtain information on the mechanisms underlying AI. Tumour samples from 26 of the familial patients were fluorescent labelled and co-hybridized with a differently labelled normal tissue DNA to metaphase chromosomes.

3.1.3 Quantitative PCR

Quantitative PCR (qPCR) experiments with TaqMan chemistry were performed on tumour tissue DNA samples from 26 familial and 26 stage and sex matched sporadic cases. The familial cases were already included in the CGH analysis. DNA was amplified using primers for microsatellite marker D20S178 and the PCR products were run on a GeneAmp 5700 Sequence detection system (AB). The relative copy numbers on chromosome 20 were determined relative to three normal tissue samples. As a reference, we used microsatellite markers D10S586, D11S1315, and D21S1904 from genomic regions stable according to the AI analysis.

3.2 Genetic analyses of AURKA (II)

3.2.1 AURKA sequencing

The AURKA coding region and exon-intron borders were sequenced in the normal tissue DNA from 10 familial CRC patients. The DNA samples were amplified and run on an ABI3730 automatic DNA sequencer according to manufacturer’s instructions (AB). The frequency of the observed sequence variants was assessed by sequencing a
larger normal tissue sample set from additional familial and sporadic CRC patients. The occurrence of the sequence variants was also assessed in Finnish cancer-free controls by sequencing.

3.2.2 Amplification analysis by sequencing and fragment analysis

After identifying AURKA sequence variants, we analyzed the possible preferential amplification of either allele in the heterozygous individuals by sequencing or fragment analysis. The PCR fragments were run on an ABI3730 automatic DNA sequencer according to manufacturer’s instructions (AB). The sequence graphs were analyzed with Chromas 2.21 (Technelysium Pty Ltd, Queensland, Australia) and the fragment analysis graphs with GeneMapper 3.0 (AB). Amplification scoring was performed according to Canzian et al. (1996).

3.3 MLH3 mutation analysis (III)

3.3.1 MLH3 sequencing and denaturing high-performance liquid chromatography (dHPLC) analysis

MLH3 sequence variants were screened in the normal tissue DNA of 30 CRC patients by sequencing the coding region and exon-intron borders. The PCR fragments were run on an ABI3100 according to manufacturer’s instructions (AB). The possible occurrence of the detected sequence variants was assessed in Finnish cancer-free controls by dHPLC. For dHPLC, the PCR fragments were denatured and reannealed, and the heteroduplex analysis was performed using automated HPLC instrumentation with an Agilent 2G experimental dsDNA 2.1 X 75 mm 3.5μ column (Agilent, Palo Alto, CA). Samples displaying aberrant dHPLC graphs were sequenced to verify the alterations.

The previously published MLH3 germline alterations (Wu et al. 2001) were also screened in the Finnish cancer-free controls by dHPLC.

3.4 Genetic analyses of SEMG1 and intergenic T9 repeats (IV)

3.4.1 SEMG1 sequencing and denaturing high-performance liquid chromatography (dHPLC) analysis

The occurrence of SEMG1 IVS2+5(T)9 mutations was determined by sequencing and dHPLC. Tumour samples from 121 MSI and 21 MSS CRC patients were analyzed by sequencing and a set of 93 MSI and 230 MSS CRCs were analyzed by dHPLC. Some of the tumour samples were analyzed using both methods to confirm their SEMG1 status. Altogether 146 MSI and 238 MSS tumour samples were analyzed. As controls, we assessed the normal tissue DNA of 144 MSI CRC patients and samples from 50 Finnish cancer-free controls by sequencing.

The coding region and exon-intron borders of SEMG1 were screened for other possible sequence variants by sequencing (exons 2 and 3) and dHPLC (exon 1). We
analyzed 20 MSI CRCs, 20 MSS CRCs, and 7 cell lines including 5 MSI CRC, 1 MSS CRC, and 1 MSI endometrial cancer cell line. Samples producing aberrant peaks in dHPLC were sequenced.

3.4.2 Evaluation of SEMG1 mRNA levels by quantitative PCR

The possible effect of the SEMG1 IVS2+5(T)9 mutation at the RNA level was assessed by qPCR with SYBR Green chemistry. RNA from six CRC cell lines was extracted and the generated cDNA was amplified using cDNA-specific primers. The PCR amplifications were monitored using a GeneAmp 5700 Sequence detection system (AB). The SEMG1 expression levels were determined relative to the GAPDH expression levels using the formula $2^{-\Delta \Delta Ct}$ (ABI PRISM 7700 Sequence Detection System User Bulletin #2, AB).

3.4.3 Evaluation of SEMG1 protein levels by Western blotting

The putative effect of the SEMG1 IVS2+5(T)9 mutation at the protein level was assessed by Western blotting. Total protein was extracted from cell lysates and cell culture media of six CRC cell lines. From each cell line, equal amounts of total protein were loaded into a 10% Tris-HCl gel. Total protein was blotted on a polyvinyl membrane and SEMG1 was detected using a polyclonal SEMG1 antibody (Biocarta, San Diego, CA). Total protein staining and the GAPDH protein levels were used as running controls. The protein band intensities were measured using a FluorChem8800 imaging system (Alpha Innotech, San Leandro, CA).

3.4.4 Sequencing of intergenic T9 repeats

The general background mutation frequency of MSI CRCs was assessed by sequencing a panel of 29 MSI CRCs for mutations in ten intergenic T9 repeats. To exclude polymorphisms, the respective normal tissue samples of patients harbouring mutations in the T9 repeats were sequenced.
RESULTS

1. Common regions of allelic imbalance (I)

After analyzing the 29 familial and 75 sporadic Finnish CRC cases with a genome-wide panel of 372 microsatellite markers, the AI frequencies for each marker in familial and sporadic patient groups were calculated. Markers indicating an association between familial cancer and AI were chosen for further analysis. Namely, 10 markers from separate genomic areas were chosen, 7 of which showed at least 10% units more AI in the familial group than in the sporadic group and the adjacent marker showed such a difference as well. The 7 markers selected from each cluster were the ones showing the biggest difference in the AI frequencies between familial and sporadic groups. Furthermore, 3 markers were chosen that harboured near 30% units more AI in the familial group although the adjacent markers showed equal AI frequencies in both groups.

In the second round of analysis, new samples from 59 familial and 64 sporadic cases were analyzed with the 10 markers. Combining results from the two rounds indicated an association between AI on chromosome 20 and familial cancer. The remaining available familial cases (n=11) and a set of sporadic cases (n=47) were then analyzed with the chromosome 20 markers. Altogether, the 99 familial and 186 sporadic cases indicated a significant difference (p < 0.01) in the AI frequencies between familial and sporadic patients on 20q13 marker D20S178.

Subsequently, the frequency of AI on D20S178 was determined in sample sets from other populations. The small Danish set of 18 familial and 24 sporadic cases showed a significant difference (p=0.002) between the two groups. However, AI was observed equally frequently in the groups consisting of 67 familial and 96 sporadic cases from the UK (p=0.68).

Based on CGH analysis performed on 26 familial CRCs, amplifications rather than deletions were underlying AI on chromosome 20q. Amplifications on 20q were found in 85% of the tumours. qPCR experiments on 26 familial and 26 sporadic cases indicated that amplifications were equally common in familial (72%) and in sporadic (73%) CRCs.

2. Analyses of AURKA (II)

We sequenced the coding region and exon-intron borders of AURKA, a candidate CRC susceptibility gene on chromosome 20q13. Sequencing of normal tissue DNA samples from 10 Finnish familial CRC patients revealed 3 missense variants (Phe31Ile, Val57Ile, and Met373Val) and a 61bp duplication in intron 4 (IVS4+2-62dup). To determine the allele frequencies of these sequence variants, we analyzed altogether 125 familial and 110 sporadic samples for the occurrence of Phe31Ile, Val57Ile, and IVS4+2-62dup variants. The Met373Val variant was screened in 42 familial and 46 sporadic cases. The allele frequencies of each sequence alteration
were similar in the familial and sporadic patient groups, as well as in the 94 cancer-free controls (Table 1).

Table 1. AURKA allele frequencies in the Finnish CRC patients and cancer-free controls.

<table>
<thead>
<tr>
<th>Sequence alteration</th>
<th>Allele frequency % (no. of cases)</th>
<th>Familial CRCs</th>
<th>Sporadic CRCs</th>
<th>Cancer-free controls</th>
</tr>
</thead>
<tbody>
<tr>
<td>IVS4+2-62dup (++/+−/−−)</td>
<td>18/82 (3/39/83)</td>
<td>17/83 (2/33/75)</td>
<td>20/80 (6/26/62)</td>
<td></td>
</tr>
<tr>
<td>91T&gt;A, Phe31Ile (TT/TA/AA)</td>
<td>74/26 (70/44/11)</td>
<td>70/30 (52/50/8)</td>
<td>72/28 (46/43/5)</td>
<td></td>
</tr>
<tr>
<td>169G&gt;A, Val57Ile (GG/GA/AA)</td>
<td>82/18 (83/39/3)</td>
<td>83/17 (75/33/2)</td>
<td>80/20 (62/26/6)</td>
<td></td>
</tr>
<tr>
<td>1683A&gt;G, Met373Val (AA/AG/GG)</td>
<td>91/9 (34/7/0)</td>
<td>95/5 (40/4/0)</td>
<td>90/10 (76/18/0)</td>
<td></td>
</tr>
</tbody>
</table>

The allele specific amplification patterns of the four sequence alterations were assessed in the heterozygous individuals. The amplification frequencies of 91T>A (Phe31Ile) alleles were significantly different from the expected (p=0.03) (Table 2), whereas the other variants did not show allele specific amplification patterns. The amplification of 91A was more pronounced in the familial group (p=0.03) than in the sporadic group (p=0.36) (Table 2). Subsequently, the clinicopathological characteristics of the familial and sporadic CRC patients were compared with respect to the genotypes at each variant. The only correlation observed was a trend between younger age at diagnosis in the group consisting of familial heterozygous carriers and homozygous individuals for the more rare 91A allele (p=0.06).

Table 2. Frequencies of AURKA 91T>A genotypes in the Finnish familial and sporadic CRC cases. The proportion of amplified alleles in the tumour samples is depicted.

<table>
<thead>
<tr>
<th>Samples analyzed (no. of samples)</th>
<th>Genotype frequencies TT/TA/AA</th>
<th>A amplified</th>
<th>T amplified</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td>Familial CRCs (125)</td>
<td>70/44/11</td>
<td>20/44 (45%)</td>
<td>5/44 (11%)</td>
<td>0.03</td>
</tr>
<tr>
<td>Sporadic CRCs (110)</td>
<td>52/50/8</td>
<td>18/50 (36%)</td>
<td>11/50 (22%)</td>
<td>0.36</td>
</tr>
<tr>
<td>Combined (235)</td>
<td>122/94/19</td>
<td>38/94 (40%)</td>
<td>16/94 (17%)</td>
<td>0.03</td>
</tr>
</tbody>
</table>

We assessed aneuploidy of the tumours by combining results from studies I and II. Aneuploidy was scored in a given chromosome arm if AI was detected in at least 50% of the informative markers. In chromosomes 9q, 11q, 13q, and 20q, more AI was observed in tumours from 91T>A heterozygous than 91T homozygous individuals with p-values of 0.04, 0.02, 0.01, and 0.007, respectively. However, AI on 2q was observed more frequently in tumours from 91T homozygous individuals with a p-value of 0.01.
3. **MLH3 mutation analysis (III)**

Sequencing of the *MLH3* coding region and exon-intron borders in the normal tissue DNA of 30 Finnish CRC patients revealed 5 missense variants: Phe390Ile, Val420Ile, Glu624Gln, Met809Val, and Asp1105Glu. The allele frequency of each variant was 1.7% (Table 3). Variants Phe390Ile, Val420Ile, Glu624Gln, and Asp1105Glu were detected in the Finnish cancer-free controls with allele frequencies of 0.6%, 1.1%, 1.1%, and 0.2%, respectively. The Met809Val variant found in a familial case with mild MSI was not found in the 713 cancer-free controls analyzed by dHPLC (Table 3). No LOH was observed in the tumour tissue of the patient harbouring the Met809Val change.

In the cancer-free control samples, 3 missense variants Asp385His, Thr942Ile, and Ser966Pro were detected with allele frequencies of 0.6%, 0.9%, and 0.9%, respectively (Table 3). These variants were not identified in the CRC patients’ normal tissue DNA.

<table>
<thead>
<tr>
<th><em>MLH3</em> variant</th>
<th>No. of CRC patients (allele frequency)</th>
<th>No. of cancer-free controls (allele frequency)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Asp385His</td>
<td>0/30</td>
<td>1/90 (0.6%)</td>
</tr>
<tr>
<td>Phe390Ile</td>
<td>1/30 (1.7%)</td>
<td>1/90 (0.6%)</td>
</tr>
<tr>
<td>Val420Ile</td>
<td>1/30 (1.7%)</td>
<td>2/90 (1.1%)</td>
</tr>
<tr>
<td>Glu624Gln</td>
<td>1/30 (1.7%)</td>
<td>4/180 (1.1%)</td>
</tr>
<tr>
<td>Met809Val</td>
<td>1/30 (1.7%)</td>
<td>0/713</td>
</tr>
<tr>
<td>Thr942Ile</td>
<td>0/30</td>
<td>3/174 (0.9%)</td>
</tr>
<tr>
<td>Ser966Pro</td>
<td>0/30</td>
<td>3/174 (0.9%)</td>
</tr>
<tr>
<td>Asp1105Glu</td>
<td>1/30 (1.7%)</td>
<td>1/268 (0.2%)</td>
</tr>
</tbody>
</table>

The previously published 10 *MLH3* variants were also screened in the cancer-free controls with an emphasis on the frameshift mutation 2578delA (Wu et al. 2001). One of the published variants, Glu624Gln, was observed in one normal tissue sample from a Finnish CRC patient and was also found in the cancer-free controls with an allele frequency of 1.1%. The other sequence variants detected by Wu et al. (2001) were not detected in the Finnish cancer-free controls. A total of 707 control samples were analyzed for the 2578delA alteration.

4. **Genetic analyses of SEMG1 and intergenic T9 repeats (IV)**

*SEMG1* intron 2 T9 repeat [IVS2+5(T)9] mutations were observed in 51% (75/146) of the MSI tumours analyzed. None of the successfully analyzed respective normal tissue samples (n=141) displayed the mutations. Furthermore, no alterations were detected in the 238 MSS tumours or in the 50 cancer-free controls. In the 31 cell lines studied, only MSI cell lines were affected with a mutation frequency of 62% (8/13) (Table 4).
A possible second hit was screened by sequencing the *SEMG1* coding region and exon-intron borders in 20 MSI CRCs, 20 MSS CRCs, and 7 cell lines including 5 MSI CRC, 1 MSS CRC, and 1 MSI endometrial cancer cell line. We observed 5 sequence variants C1386T, Arg447His, Arg457Gln, C1417T, and IVS2+30A>G that were present also in the respective normal tissue DNA. The sequence alterations were also found in the cancer-free controls with similar allele frequencies.

The putative effects of the IVS2+5(T)9 mutations were analyzed at the RNA and protein levels. No differences were detected in the RNA or protein levels or sizes between the six CRC cell lines with different *SEMG1* IVS2+5(T)9 status.

The clinicopathological characteristics of the MSI CRC patients with (n=75) and without (n=71) the IVS2+5(T)9 mutations were compared. No correlations were observed between the mutation status and age at onset, sex distribution, location or staging of the tumour, or presence of germline MMR mutation.

Subsequently, the general background mutation frequency of MSI CRCs was assessed by sequencing 29 MSI CRCs for mutations in ten intergenic T9 repeats. Polymorphisms were excluded by screening the respective normal tissue samples from those patients displaying mutations in the tumour samples. After analyzing the tumour and normal tissue samples, we detected mutation frequencies of 0% for AL161657, 7% for AC103870, 10% for AL445240, 10% for AC108706, 27% for AP002759, 28% for AP002801, 29% for AC008163/1, 30% for AC008163/2, 27% for AC004006, and 86% for AC027013. Due to the high mutation frequency in the T9 repeat in AC027013, we extended our analysis to include a total of 145 MSI CRCs and respective normal tissue samples. Mutation frequency of 70% (102/145) was observed in the tumour samples, whereas none of the 143 available and successfully analyzed normal tissue samples displayed the mutation (Table 4).

### Table 4. Frequency of *SEMG1* and AC027013 T9 mutations.

<table>
<thead>
<tr>
<th>Samples analyzed</th>
<th>Mutation frequency, % (no. of cases)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td><em>SEMG1</em></td>
</tr>
<tr>
<td>MSI CRCs</td>
<td>51 (75/146)</td>
</tr>
<tr>
<td>MSS CRCs</td>
<td>0 (0/238)</td>
</tr>
<tr>
<td>Cancer cell lines</td>
<td></td>
</tr>
<tr>
<td>MSI CRC</td>
<td>70 (7/10)</td>
</tr>
<tr>
<td>MSS CRC</td>
<td>0 (0/18)</td>
</tr>
<tr>
<td>MSI prostate</td>
<td>0 (0/2)</td>
</tr>
<tr>
<td>MSI endometrial</td>
<td>100 (1/1)</td>
</tr>
<tr>
<td>Normal tissue samples</td>
<td>0 (0/141)</td>
</tr>
<tr>
<td>Cancer-free controls</td>
<td>0 (0/50)</td>
</tr>
</tbody>
</table>

nd=not determined
DISCUSSION

1. Novel colorectal cancer susceptibility loci (I)

The known polyposis and non-polyposis CRC syndromes together account for approximately 5% of all CRC cases. Since an estimated 35% of all CRCs can be attributed to heritable factors, the existence of novel susceptibility genes is evident (Lichtenstein et al. 2000). Linkage studies performed in the past decade further support the existence of novel genes predisposing to CRC. For example, linkage against known CRC susceptibility loci was observed in 10 families with excess clustering of the disease (Lewis et al. 1996). Furthermore, linkage to novel loci on 9q and 15q has been obtained by analyzing affected siblings or Ashkenazi families, respectively (Tomlinson et al. 1999, Jaeger et al. 2003, Wiesner et al. 2003). Most of the familial CRC clustering might be explained by a number of more prevalent low-penetrance susceptibility alleles. Implications for the general public health can be vast thus making the identification of these novel susceptibility genes important.

There are typically only two affected core family members in the non-syndromic CRC families of our sample collection. The underlying predisposing genes therefore confer a modestly increased cancer risk. Identification of these genes is best performed by means other than linkage analysis, especially since samples are mostly available only from the proband. Therefore, a genome-wide microsatellite marker analysis was performed to obtain information on commonly deleted or amplified regions, to form a basis for candidate gene analysis. This approach is based on the notion that AI can be detected more frequently in the familial cases at a given susceptibility locus. This allelotyping approach has the advantage that it does not require multiple samples from affected relatives. Although this method may be hampered by genetic heterogeneity, it is feasible in homogeneous populations. The Finnish population is ideal for this kind of studies due to founder effect, population isolates, and genetic drift that together have formed the gene pool homogeneous. Furthermore, this approach identified 13q deletions especially in familial breast cancer cases prior localizing the 13q gene BRCA2 associated with breast cancer susceptibility (Thorlacius et al. 1991, Wooster et al. 1995).

We used a genome-wide microsatellite marker panel to allelotype Finnish familial CRCs with mild clustering of cancer in the family, as well as Finnish sporadic CRCs. All samples used in this study were from patients with MSS CRC. Furthermore, known polyposis cases were excluded from the analyses. We identified a significant difference in the AI frequencies between familial and sporadic cases on chromosome 20q13. From the microsatellite marker analysis graphs, it is difficult to interpret whether deletions or amplifications are causing the difference between the allele intensities. Based on our CGH experiments, the imbalances detected on 20q13 were caused by amplifications indicating a potential oncogenic mode of action for the putative CRC susceptibility gene within the affected region.

When assessing the frequency and amplitude of amplifications in the familial and sporadic CRCs by qPCR, both familial and sporadic cases seemed to harbour amplifications equally frequently. The qPCR method utilized here cannot separate the
alleles. Therefore, both alleles of sporadic CRCs might be frequently amplified whereas amplifications seem to preferentially target the other allele in the familial cases. Similar phenomenon has been seen with preferential amplification of a mutant MET allele in papillary renal cell cancer indicating the importance of mutant MET allele copy number in cancer formation (Fischer et al. 1998).

In previous studies, 20q amplifications have been detected in sporadic CRC cases from different populations but this is the first time amplifications were seen in excess in familial cases. Subsequently, sample sets obtained from CRC patients from Denmark and UK were analyzed for imbalances on 20q13. The small Danish set gave similar results but the UK set showed no differences between the familial and sporadic groups. These results indicate the possibility of inter-population variation due to the involvement of different susceptibility genes or modifying factors (genetic or environmental) in different populations.

The p-values obtained from the AI frequency comparisons between familial and sporadic groups have not been corrected for multiple testing. It is acknowledged that the probability of finding significant correlations increases with increasing numbers of comparisons made. However, the utility of, for example Bonferroni adjustments for multiple testing can be questioned. Although it protects from the type I errors (false positives), typically type II errors increase and true associations can be missed especially if the number of tests is high.

The identification of 20q13 amplifications preferentially in the familial CRCs suggests the presence of putative oncogene(s) predisposing to CRC within the amplicon. Some candidate CRC genes reside within the amplicon. For example, AURKA has recently been suggested to function as a low-penetrance susceptibility gene (Ewart-Toland et al. 2003) and ZNF217 copy number changes have been linked to poorer survival and metastatic potential (Hidaka et al. 2000, Rooney et al. 2004). Furthermore, the potential tumorigenic role of other less well characterized genes within the 20q13 amplicon cannot be overlooked. Identification of the causative gene alterations would benefit from a more detailed delineation of the amplified region by using for example a more dense microsatellite marker map or array-CGH. Furthermore, expression analysis of genes within the amplicon and discovery of sequence alterations within candidate genes followed by association analyses might facilitate the identification of causative alterations.

2. AURKA, a novel low-penetrance susceptibility gene (II)

In study I, allele specific amplifications of 20q13 were seen more frequently in familial than in sporadic CRC cases indicating the possible existence of novel susceptibility gene(s) within the amplicon. A recently published association analysis suggested a role for 20q13 gene AURKA in CRC predisposition (Ewart-Toland et al. 2003). AURKA located within a genomic region orthologous to a mouse quantitative trait locus (QTL) associated with cancer susceptibility. Analyses on unselected CRC and healthy control individuals of Northern European ancestry indicated a trend between 91T>A (Phe31Ile) change and higher CRC risk. The potential tumorigenic role of the sequence alteration was highlighted by a stronger transforming potential of the 91A allele in cell culture experiments. Furthermore, a preferential amplification of
91A allele was detected together with a more aneuploid phenotype of the tumours from heterozygous individuals. The 91T>A change could therefore affect the normal function of AURKA in chromosome segregation (Zhou et al. 1998, Ewart-Toland et al. 2003). AURKA is also involved in the p53 pathway by controlling the degradation of p53 together with MDM2 (Katayama et al. 2004). AURKA overexpression leads to an increased degradation of p53 which in turn leads to downregulation of cell cycle checkpoint pathways and oncogenic transformation of affected cells (Katayama et al. 2004).

Since the study of Ewart-Toland et al. (2003), association analyses on other cancer types have produced similar results emphasizing the role of AURKA as a low-penetration susceptibility gene (Dai et al. 2004, DiCioccio et al. 2004, Miao et al. 2004, Sun et al. 2004). The results have been somewhat conflicting as to whether the 91T>A change contributes solely or in combination with other AURKA changes. However, a recently published meta-analysis on several cancer types revealed a significantly increased cancer risk in 91T>A heterozygous individuals (Ewart-Toland et al. 2005).

In study II, the presence of possible AURKA sequence changes and their preferential amplification patterns were evaluated in samples from CRC patients with and without family history of the disease. The samples used in this study were from patients with non-syndromic MSS CRC. Prior to this study, confirmation to the preferential amplification of 91A had not been pursued and the role of AURKA had not been evaluated in familial cancer.

Sequencing of the AURKA coding region and exon-intron borders revealed 3 missense changes (Phe31Ile, Val57Ile, and Met373Val) and a 61bp duplication in intron 4 (IVS4+2-62dup). The frequency of these alterations was assessed in a larger set of normal tissue samples from familial and sporadic patients and the amplification patterns were evaluated in the heterozygous individuals. The frequency of each sequence variant was similar in both familial and sporadic groups as well as in cancer-free controls. The amplification patterns of 169G>A (Val57Ile), 1683A>G (Met373Val), and IVS4+2-62dup did not differ from the expected. Consistent with previous findings, the 91A allele was more frequently amplified than the 91T allele in the heterozygous individuals (p=0.03). As a novel finding, a difference between the familial and sporadic groups was detected: in the familial group, the amplification of 91A allele was significantly more frequent (p=0.03) whereas in the sporadic group a non-significant p-value of 0.36 was obtained.

Patients with different AURKA genotypes were compared within the familial and sporadic groups to find possible genotype-phenotype correlations. Comparisons were carried out between patients homozygous for the common allele and patient group containing both heterozygous carriers and homozygous patients for the rarer allele. No indications were obtained for the tumorigenic role of Val57Ile, Met373Val, or IVS4+2-62dup changes from comparisons of age at diagnosis, sex, tumour location, or staging of the tumour. However, familial 91T>A heterozygotes and 91A homozygotes were diagnosed at a younger age than familial patients with the more common 91T genotype (p=0.06). This trend was not seen in the sporadic group.
Ewart-Toland et al. (2003) further observed that tumours from 91T>A heterozygous individuals were more aneuploid, a finding that has not been confirmed to date. We combined the results from studies I and II to evaluate the ploidy of each tumour with respect to the AURKA genotype. We obtained evidence in favour of more frequent chromosomal imbalances on 9q, 11q, 13q, and 20q in heterozygous individuals. However, chromosome 2q was more aneuploid in tumours from 91T homozygous individuals.

In study II we sought for a potential target gene for the 20q13 amplification observed in study I. We evaluated the most promising candidate AURKA but, understandably, other possible amplification targets may exist. Other putative candidates, such as ZNF217 could be analyzed to further elucidate the implications of the observed amplification. However, the results produced in study II indicate that the 20q13 amplifications target AURKA in an allele-specific manner and support the role of the 91T>A change in CRC predisposition. The difference in the amplification patterns of AURKA 91T and 91A alleles between the familial and sporadic groups indicates that the 91A allele could have a more pronounced effect in cases with family history of the disease. A modifying genetic factor could further emphasize the effect of 91A change in the familial patients. Further analyses are thus warranted.

3. Role of MLH3 in microsatellite-unstable colorectal cancer (III)

Mutations in the MMR genes MLH1 and MSH2 have been shown to underlie a majority of the HNPCC cases whereas a smaller proportion is due to mutations in MMR genes PMS2 and MSH6 (Peltomäki and Vasen 2004). Causative mutations have not been detected in a subset of the families meeting the Amsterdam criteria, indicating the possible involvement of additional gene(s) in MSI CRC (Liu et al. 1996). In 2000, the identification of a human homolog of the S. cerevisiae Mlh3p was published (Lipkin et al. 2000). Mlh3p had previously been shown to function in MMR in yeast (Flores-Rozas and Kolodner 1998). Lipkin et al. (2000) suggested interactions between human MLH3 and MLH1 and, furthermore, a link between MLH3 defects and MSI. These findings prompted several genetic analyses on the role of MLH3 in CRC predisposition. The first studies revealed no germline mutations or reported missense type of changes that did not present themselves as attractive disease-causing candidates (Loukola et al. 2000, Lipkin et al. 2001).

In 2001, Wu and co-workers suggested that MLH3 could have a more pronounced role in HNPCC than had previously been considered (Wu et al. 2001). They identified 10 germline MLH3 variants, including a one bp deletion and 9 missense variants, that were absent in the cancer-free controls. All variants were detected in patients suspected of HNPCC. Segregation analyses of the observed variants could not be performed. MSI analysis of tumours from patients with MLH3 variants indicated more pronounced MSI in dinucleotide and tetranucleotide repeats than in mononucleotide repeats, atypical of MSI tumours caused by mutations in other MMR genes (Wu et al. 2001).

To further assess the role of MLH3 in CRC predisposition, we selected CRC cases for germline mutation analysis based on the MLH3 defect features suggested by Wu et al. (2001). MLH3 germline mutations were screened by sequencing the coding region
and exon-intron borders of all available mutation negative HNPCC cases (n=7), as well as familial (n=17) and sporadic (n=6) CRC patients with more pronounced MSI in dinucleotide and tetranucleotide repeats than in mononucleotide repeats.

We identified five missense mutations (Phe390Ile, Val420Ile, Glu624Gln, Met809Val, and Asp1105Glu) with allele frequencies of 1.7%. Of note, the Glu624Gln change had been previously published by Wu et al. (2001) as a potential CRC-associated mutation. We detected all variants except Met809Val also in the cancer-free controls with similar allele frequencies as in CRC patients. The Met809Val change was not found in 713 cancer-free controls analyzed. However, it seems to affect a non-conserved residue located outside functionally relevant domains (Kondo et al. 2001).

In study III, 3 missense variants (Asp385His, Thr942Ile, and Ser966Pro) were detected in the Finnish cancer-free controls with allele frequencies varying between 0.6 and 0.9%. These variants were not present in the CRC cases analyzed. The cancer-free controls were analyzed for the occurrence of previously identified MLH3 variants with a main focus on the most promising disease-associated mutation, 2578delA (Wu et al. 2001). One of the previously identified missense variants (Glu624Gln) was found in the controls. The other sequence changes, including the 2578delA, were not detected in the controls or in CRC cases analyzed.

Recently de Jong et al. (2004) examined the allele frequencies of a subset of the MLH3 exon 1 variants published by Wu et al. (2001). The frequencies of the variants were determined in sporadic CRC patients and controls. A Ser817Gly missense variant reported by Wu et al. (2001) with an allele frequency of 0.2% was identified in the sporadic CRC patients and controls with similar allele frequencies (0.3% and 0.2%, respectively). Another missense variant affecting the same amino acid residue (Ser817Arg) was identified by de Jong et al. (2004) in the sporadic CRC patients. This variant was absent in the cancer-free controls. The truncating 2578delA mutation reported by Wu et al. (2001) was not found in 467 sporadic CRC patients or 497 controls. Furthermore, the Met809Val missense variant identified in study III was found in the Dutch sporadic CRC patients and control individuals with allele frequencies of 0.2% and 0.3%, respectively (de Jong et al. 2004).

Liu et al. (2003) have also analyzed the occurrence of MLH3 variants in 70 unrelated individuals from families suggestive of a genetic predisposition to CRC. They identified one frameshift mutation (885delG) and 11 missense mutations in 16 index patients with MSI-negative tumours. The results therefore suggest mechanisms other than deficient MMR for MLH3-associated CRC. Liu and co-workers (2003) are the first to study the segregation of detected MLH3 variants in the cancer families. Segregation analyses supported the role of 3 MLH3 variants (885delG, G2221T, and T3826C) as causative for the disease, but most probably representing low risk alleles. They hypothesized that MLH3 could contribute to CRC risk together with other low-penetrance genes (Liu et al. 2003).

Taken together, MLH3 seems to harbour a wide range of sequence alterations. Most of the variants seem to lack apparent functional significance because of their localization in non-conserved residues, lack of clear segregation, and similar allele frequencies in familial and sporadic cancer cases as well as in controls (Loukola et al. 2003).
2000, Lipkin et al. 2001, Wu et al. 2001, Liu et al. 2003, de Jong et al. 2004, study III). The role of MLH3 as a low-penetrance susceptibility gene needs to be studied further by association analyses. Since there seems to be inter-population differences in the MLH3 variants detected, population specific association analyses would have to be performed.

4. High frequency of microsatellite mutations in SEMG1 and intergenic T9 repeats (IV)

Mutations in the MMR genes lead to general instability phenomenon characterized by the accumulation of microsatellite mutations. These mutations affect tumour growth when occurring within genes having growth related or caretaker functions. A number of genes have been studied for MSI and mutations have been detected in the coding region of genes coding for growth factors, growth factor receptors, transcription factors, and apoptotic factors (Woerner et al. 2003). In 1998, Boland et al. suggested criteria that could be used to ascertain real MSI target genes. Two of the criteria, high mutation frequency and functional evidence, have been the most widely accepted.

Most of the MSI target gene studies published to date have relied solely on mutation frequency data in defining true target genes, partly because obtaining and interpreting functional evidence can be a laborious task. In the absence of functional evidence, the importance of control samples accentuates. To avoid miscalling of mutations, normal tissue DNA samples and MSS samples need to be analyzed. Furthermore, to obtain more reliable estimation on the true mutation frequency, primary tumours need to be assessed since estimations based on MSI cell lines can lead to overestimations (Duval et al. 2001, Suzuki et al. 2002). To facilitate true target gene identification, mutation frequency cut-off levels have been suggested (Duval et al. 2001, Woerner et al. 2003). However, it has also been presented that the tumorigenic role of a given gene can be difficult to ascertain solely from the mutation frequency data (Duval and Hamelin 2002, Perucho 2003). For example, mutations in survivor genes with fundamental roles in cellular homeostasis can be negatively selected for. In addition, mutations in co-operator genes can be sparse and have a mild but nonetheless significant effect that is further emphasized when occurring together with mutations in other genes operating in the same pathway (Yamamoto et al. 2000).

Generally non-coding microsatellite mutations are thought to be inconsequential background events and the mutation frequencies are therefore considered to be low. However, the functional importance of repeat elements outside the coding regions is not fully understood. They have been suggested to play a role in genome evolution, chromatin organization, and in the regulation of gene activity and DNA metabolic processes (Kashi et al. 1997, Li et al. 2002). In fact, some non-coding microsatellite mutations have been shown to have functional consequences. For example, 93% of primary MSI CRCs were seen to harbour mutations in MRE11 intron 4 acceptor site that lead to a partial splicing defect and reduced mRNA and protein levels (Giannini et al. 2002). Furthermore, the ability of MRE11 to function in S-phase checkpoint was impaired (Giannini et al. 2002). In addition, evidence for the role of CA-repeats and CA-rich elements as splicing regulators has recently been obtained (Hui et al. 2003, Hui et al. 2005). CA-elements were seen to control splicing by acting either as enhancers or silencers depending on the distance between the elements and the
alternative 5’ splice sites (Hui et al. 2005). In another recently published study, CA-repeats were seen to control the promoter activity of Cyr61 (Wang et al. 2005). These studies highlight the potential importance of non-coding microsatellites as, for example, gene expression regulators and indicate the need for further, large-scale studies of non-coding microsatellite mutations and their implications.

When assessing the background mutation frequency of MSI CRCs, we observed a relatively high mutation frequency in a T9 repeat in SEMG1 intron 2. Subsequently we extended the analysis to include a larger number of MSI CRCs and a mutation frequency of 51% was observed. No SEMG1 T9 repeat alterations were detected in the MSS CRCs or respective normal tissue controls analyzed, thus the SEMG1 mutations seem to associate with MMR deficiency. The mutation frequency observed here is much higher than the cut-off levels used to determine true MSI target genes (Duval et al. 2001, Woerner et al. 2003) and much higher than in some of the proposed true target genes. To search for a possible effect of the SEMG1 intron 2 mutations, we conducted RNA and protein level analyses but observed no differences between CRC cell lines with different SEMG1 status. This indicates either that the mutation has an effect that cannot be detected by the methods used here, or that the mutations are bystander events without functional consequence. If the latter hypothesis holds, it shows that extremely high mutation frequencies can be detected in non-coding microsatellites even without selection pressure.

To assess the background mutation frequency of MSI CRCs further, we analyzed a panel of 29 primary MSI CRCs for mutations in 10 intergenic T9 repeats. The repeats were chosen from genomic regions distant from known genes or ESTs. The T9 repeats were altered in 0 – 86% of the studied CRCs. To exclude germline polymorphisms, we analyzed the normal tissue DNA of patients showing alterations in the tumour sample. Only one of the repeats showed alterations in the normal tissue samples. However, the 86% mutation frequency in a T9 repeat in genomic contig AC027013 persisted.

Subsequently, we extended our analysis to the whole set of 145 MSI CRCs and observed a mutation frequency of 71% without mutations in the respective normal tissue DNA samples. This result further indicates that high mutation frequencies can be found without any apparent functional consequence of the mutations. Of note, after scrutinizing the latest Celera database update, two novel transcripts were detected in the AC027013 T9 repeat region. With respect to the first transcript, the T9 repeat locates in the first intron, 2.7 and 16.9kb away from the exon-intron borders. The other transcript was located downstream of the T9 repeat with 1.1kb between the repeat and transcription start site: putative promoter regions were sought by several programs but none were found. Thus these observations speak against a functional consequence of the T9 repeat mutations. Due to the constantly growing knowledge on human transcriptomics, the intergenic nature of the T9 repeats studied here will need to be re-evaluated. To date, new transcripts have not been presented in the vicinity of AC027013 T9 repeat.

The variability observed in the mutation frequencies between non-coding repeats of the same composition and size might be at least partly explained by mechanisms controlling DNA replication. Replication accuracy is controlled by DNA polymerase fidelity as well as DNA sequence context. Polymerase fidelity is dependent on the
geometric selection of the correct nucleotide for insertion into DNA as well as on the 3’-5’ exonuclease activity of the polymerase (Goodman 1997). Mechanisms underlying DNA context-dependent replication accuracy are less well understood but they may depend on DNA duplex stability and base-stacking properties. Furthermore, GC content or point substitution rates of flanking sequences have been suggested to affect microsatellite stability (Brock et al. 1999, Santibanez-Koref et al. 2001).

Taken together, the results provided by study IV emphasize the need of thorough large-scale analyses of microsatellite mutations in non-coding regions of the genome. Information on intronic and intergenic repeats from several genomic regions are needed since the mutation frequencies seem to vary quite extensively between repeats of the same base composition and size. Results produced by these studies would benefit the ongoing MSI target gene analyses and provide material for sequence comparisons to identify sequence elements controlling replication fidelity. As indicated by the results in study IV, mutation frequency alone cannot be utilized to identify putative MSI target genes. Thus, the tumorigenic role of the previously published candidate MSI target genes needs to be interpreted with caution in the absence of functional evidence.
CONCLUSIONS AND FUTURE PROSPECTS

Many colorectal cancer (CRC) predisposition syndromes have been characterized to date and the molecular changes driving CRC progression are perhaps the best characterized of all human cancer types. However, many questions remain and accumulating evidence suggests the existence of novel CRC susceptibility genes, especially of low-penetrance. This study aimed at identifying novel genes and loci associated with CRC predisposition and progression in both chromosomal instability (CIN) and microsatellite instability (MSI) pathways.

1. Identification of novel CRC susceptibility loci was pursued by assessing common regions of allelic imbalance in CIN CRCs. Chromosome 20q13 was seen to harbour amplifications more frequently in cases with family history of the disease, indicating the potential existence of novel predisposing oncogene(s) within the amplicon. Further investigations on the importance of the amplification are warranted. The putative tumorigenic role of one 20q13 candidate gene, *AURKA*, was analyzed in study II.

2. A candidate CRC susceptibility gene on chromosome 20q13, *AURKA*, was assessed in CIN CRC cases with and without family history of the disease. The observed preferential amplification of 91A allele in the familial cases and association with younger age at diagnosis support the role of *AURKA* as a novel CRC susceptibility gene. Further analyses could elucidate the potential modifying factors of *AURKA*-associated CRC. Studies I and II indicate the feasibility of using chromosomal clues in the identification of candidate loci and causative cancer genes.

3. A subset of MSI CRCs cannot be ascribed to the MMR genes currently associated with CRC predisposition. The role of MMR gene *MLH3* was assessed in mutation negative HNPCC cases, as well as in familial and sporadic CRC patients displaying distinct MSI features. Five missense mutations were identified, four of which were found in cancer-free controls as well. The remaining missense mutation is not an attractive disease-causing candidate. The results provided by study III and other recently published *MLH3* studies suggest a minor role for *MLH3* in CRC predisposition, to be evaluated further by association analyses.

4. MSI target genes have previously been proposed mainly based on high mutation frequency data. In study IV, two non-coding microsatellites were seen to display mutation frequencies higher than those observed in the coding regions of many suggested MSI target genes. Thus, mutation data alone is not sufficient to indicate the role of a given gene in MSI tumorigenesis and functional evidence should therefore be provided. MSI target gene studies would benefit from evaluations on the general background mutation frequency of MSI CRCs. Furthermore, information provided therein could elucidate the mechanisms governing the sequence context dependence of DNA replication fidelity.
ACKNOWLEDGEMENTS

This study was carried out at the Department of Medical Genetics, Haartman Institute and Biomedicum Helsinki, University of Helsinki, during 2001-2005. The former and present heads of the Department of Medical Genetics, Pertti Aula, Anna-Elina Lehesjoki, Leena Palotie, Kristiina Aittomäki, and Päivi Peltonäki are thanked for providing excellent research facilities.

I am sincerely grateful to my supervisor Lauri Aaltonen for his support and guidance during these years and for giving me the opportunity to work in his excellent research group. The atmosphere in the lab has been truly remarkable and inspiring thanks to Lauri. I believe that Lauri’s enthusiasm and innovative approach has taught a great deal to all of us that have had the pleasure to work in his team.

I am also very grateful to my other supervisor Auli Karhu for her excellent and never-ending support as well as for her valuable ideas during the course of this work. Furthermore, I wish to thank Auli for the discussions we have had - both scientific and otherwise - and for her great sense of humour that always kept me going.

I thank the reviewers of my thesis, professor Raquel Seruca and docent Outi Monni for encouraging comments and excellent ideas on how to improve the thesis.

I thank all the co-authors and collaborators that have greatly facilitated this work: Lara Lipton, Yan Aalto, Huw Thomas, Karin Birkenkamp-Demtröder, Shirley Hodgson, Jukka-Pekka Mecklin, Heikki Järvinen, Sakari Knuitila, Sarah Halford, Torben Ørntoft, Ian Tomlinson, Richard Houlston, Pertti Sistonen, Päivi Peltonäki, Thomas Barber, Riitta Koistinen, and Bert Vogelstein.

A huge thanks to the rest of the Aaltonen Team, you have made these years very enjoyable: the PhD students Antti, Heli L, Heli S, Marianna, Pia, Päivi, Rainer, Sakari, Sanna, and Taru - a special thanks to my roommates Sanna, both Helis, and Pia for the time spent discussing various extracurricular topics and science, to Päivi for guiding me during the early days when things seemed bewildering, to Rainer for showing how to party like nothing else mattered; the postdocs Virpi and Pia - a special thanks to Virpi for her excellent support and for sharing her valuable ideas on numerous projects; the skillful lab team Anniina, Iina, Inga-Lill, and Mikko - a sincere thanks for your patient assistance with the projects that sometimes seemed endless; our pathologist Reijo for his valuable assistance with the slides; our research nurses Sini and Katja - a special thanks to Sini for her excellent help with the samples and taking care of routine business when I was busy with other things; the newcomers in the team Eeva-Mari, Miranda, Päivi, and Silva; and finally the past members Annika, Diego, Kirsi, Maija, Nina, Susa, Taija, and Ullis - a special thanks to Susa for the enjoyable and educative times during the shared projects, to Taija for inspiring discussions, and to Diego for his valuable views during the confusing times of project IV. A hearty thanks to all you guys. I will never forget these years or the numerous parties - and the pictures won't let me!
I also wish to acknowledge the personnel of the Finnish Genome Center and the sequencing and fragment analysis facilities in Biomedicum and Haartman Institute whose contribution has been of great importance during the course of this work.

A huge thanks to all my friends for the excellent times spent in various locations - for the unforgettable get-togethers, parties, sporting events, and board games that have kept my head together and balanced my life in a profound way.

And of course - a warm and sincere thanks to my amazing family for the endless support and for understanding my hectic schedule. I am truly grateful to Tomi for his love and support during these years - he has kept me on track and reminded me of the things that really matter in life. And of course my sweet Mom Sirja and Dad Pentti, my superb sisters Kirsi, Saara, and Virpi, and brothers-in-law Make, Peter, and Sami, and their lovely and energetic children Tino, Kaapo&Sofia&the new baby boy, and Nea&Niko, and last but not least my second family Ritva, Lauri, and Jani - you are all simply the best!

This study was supported by grants from the Maud Kuistila Memorial Foundation, Ida Montin Foundation, and Paulo Foundation.
REFERENCES


