Genetic predisposition to colorectal cancer in young patients and in the general population

Tomas Tanskanen

Faculty of Medicine
Department of Medical and Clinical Genetics, Medicum
Genome-Scale Biology Research Program, Research Programs Unit
Doctoral Programme in Biomedicine
University of Helsinki
Finland

Academic dissertation

To be publicly discussed, with the permission of the Faculty of Medicine, University of Helsinki, in Biomedicum Helsinki 1, Lecture Hall 3, Haartmaninkatu 8, Helsinki, on December 20th 2018, at 12 noon.

Helsinki 2018

Supervised by

Lauri A. Aaltonen, M.D., Ph.D., Academy Professor
Dept. of Medical and Clinical Genetics, Medicum
Genome-Scale Biology, Research Programs Unit
Faculty of Medicine, University of Helsinki, Finland

Sari Tuupanen, Ph.D.
Dept. of Medical and Clinical Genetics, Medicum
Genome-Scale Biology, Research Programs Unit
Faculty of Medicine, University of Helsinki, Finland

Kimmo Palin, Ph.D.
Dept. of Medical and Clinical Genetics, Medicum
Genome-Scale Biology, Research Programs Unit
Faculty of Medicine, University of Helsinki, Finland

Reviewers appointed by the Faculty

Asta Försti, Ph.D.
Department of Molecular Genetic Epidemiology,
German Cancer Research Center, Germany
Center for Primary Health Care Research,
Lund University, Sweden

Kati Kristiansson, Ph.D., Docent
Dept. of Public Health Solutions,
Genomics and Biomarkers Unit,
National Institute for Health and Welfare, Finland

Opponent appointed by the Faculty

Vesa Kataja, M.D., Ph.D., Professor
Institute of Clinical Medicine,
University of Eastern Finland, Finland
Jyväskylä Central Hospital,
Central Finland Health Care District, Finland

ISBN 978-951-51-4746-2 (paperback)
ISBN 978-951-51-4747-9 (PDF)
http://ethesis.helsinki.fi/
Unigrafia Oy, Helsinki 2018

Contents

List of original publications

The thesis is based on the following publications that are referred to in the text by the Roman numerals I-III.

I **Tanskanen T**, Gylfe AE, Katainen R, Taipale M, Renkonen-Sinisalo L, Mecklin JP, Järvinen H, Tuupanen S, Kilpivaara O, Vahteristo P, Aaltonen LA. Exome sequencing in diagnostic evaluation of colorectal cancer predisposition in young patients. *Scand J Gastroenterol* 2013;48(6):672-8. DOI: 10.3109/00365521.2013.783102.

II **Tanskanen T**, Gylfe AE, Katainen R, Taipale M, Renkonen-Sinisalo L, Järvinen H, Mecklin JP, Böhm J, Kilpivaara O, Pitkänen E, Palin K, Vahteristo P, Tuupanen S, Aaltonen LA. Systematic search for rare variants in Finnish early-onset colorectal cancer patients. *Cancer Genet* 2015;208(1-2):35-40. DOI: 10.1016/j.cancergen.2014.12.004.

III **Tanskanen T**, van den Berg L, Välimäki N, Aavikko M, Ness-Jensen E, Hveem K, Wettergren Y, Bexe Lindskog E, Tõnisson N, Metspalu A, Silander K, Orlando G, Law PJ, Tuupanen S, Gylfe AE, Hänninen U, Cajuso T, Kondelin J, Sarin AP, Pukkala E, Jousilahti P, Salomaa V, Ripatti S, Palotie A, Järvinen H, Renkonen-Sinisalo L, Lepistö A, Böhm J, Mecklin JP, Al-Tassan NA, Palles C, Martin L, Barclay E, Tenesa A, Farrington SM, Timofeeva MN, Meyer BF, Wakil SM, Campbell H, Smith CG, Idziaszczyk S, Maughan TS, Kaplan R, Kerr R, Kerr D, Buchanan DD, Win AK, Hopper J, Jenkins MA, Newcomb PA, Gallinger S, Conti D, Schumacher F, Casey G, Cheadle JP, Dunlop MG, Tomlinson IP, Houlston RS, Palin K, Aaltonen LA. Genome-wide association study and meta-analysis in Northern European populations replicate multiple colorectal cancer risk loci. *Int J Cancer* 2018;142(3):540-546. DOI: 10.1002/ijc.31076.

The publications are reproduced with the permissions of the copyright holders.

Abbreviations

CI	Confidence interval
CRC	Colorectal cancer
DNA	Deoxyribonucleic acid
FAP	Familial adenomatous polyposis
FIT	Fecal immunochemical test
gFOBT	Guaiac-based fecal occult blood test
GI	Gastrointestinal
GWAS	Genome-wide association study
IBD	Inflammatory bowel disease
IHC	Immunohistochemistry
IQR	Interquartile range
IRR	Incidence rate ratio
JP	Juvenile polyposis
LD	Linkage disequilibrium
LOD	Logarithm of the odds
LoF	Loss-of-function
LOH	Loss of heterozygosity
LS	Lynch syndrome
MAF	Minor allele frequency
MAP	*MUTYH*-associated polyposis
MEN	Multiple endocrine neoplasia
MLPA	Multiplex ligation-dependent probe amplification
MMR	Mismatch repair
MSI	Microsatellite instability
MSS	Microsatellite stability
NGS	Next-generation sequencing
OR	Odds ratio
PCR	Polymerase chain reaction
PHTS	*PTEN* hamartoma tumor syndrome
PJS	Peutz-Jeghers syndrome
pLI	Probability of loss-of-function intolerance
PRS	Polygenic risk score
RCT	Randomized clinical trial
RNA	Ribonucleic acid
RR	Relative risk
SNP	Single-nucleotide polymorphism
TGF	Transforming growth factor
TNM	Tumor-node-metastasis
TSG	Tumor suppressor gene
VUS	Variant of uncertain significance
WES	Whole-exome sequencing
WGS	Whole-genome sequencing
XLA	X-linked agammaglobulinemia

Abstract

Nearly one third of people in developed countries will be diagnosed with cancer in their lifetime. Colorectal cancer (CRC) is the third most common cancer worldwide and accounts for 10% of all new cancers. It is considered one of the most preventable cancers and is often curable if diagnosed early. The risk of developing CRC increases with age and is influenced by both hereditary and environmental factors. Population screening reduces morbidity and mortality from the disease, but persons with genetic predisposition may benefit from more intensive screening and other risk-reducing interventions. However, most of the heritability of CRC remains unexplained. Therefore, more research is needed to define the genetic architecture of CRC susceptibility and to develop strategies to identify individuals with substantial genetic risk.

The first aim of this thesis was to study the diagnostic approach to hereditary cancer syndromes in patients with early-onset CRC. We investigated a series of 38 CRC patients diagnosed before age 40 years, 15 of whom had been diagnosed with hereditary CRC syndromes. To assess the practical feasibility and added value of whole-exome sequencing (WES) as a diagnostic test, we performed WES on 23 early-onset CRC patients with unknown etiology. Ten high-penetrance CRC predisposition genes (*MLH1*, *MSH2*, *MSH6*, *PMS2*, *APC*, *MUTYH*, *SMAD4*, *BMPR1A*, *STK11* and *PTEN*) were analyzed for nonsynonymous variants, and family histories were acquired from national population registries. Hereditary CRC syndromes were diagnosed in 42% (16/38; 95% confidence interval (CI), 26%-59%) of the early-onset CRC patients, including 12 patients (32%) with Lynch syndrome (LS), three patients (7.9%) with familial adenomatous polyposis (FAP) and one patient (2.6%) with juvenile polyposis (JP). WES revealed only one additional pathogenic germline variant in *MLH1*. The majority of non-syndromic patients had negative family history and microsatellite-stable (MSS) CRC. Although the prevalence of hereditary CRC syndromes was high in this population, the diagnostic yield of WES was not superior to microsatellite instability (MSI) testing and clinical assessment for gastrointestinal (GI) polyposis.

The second aim was to study the contribution of rare germline variants to early-onset CRC. Rare protein-coding variants may have clinical impact and can directly implicate genes that are involved in the pathogenesis of CRC. Also, studies in isolated populations such as Finland may be useful in the identification of rare disease-causing variants. To identify candidate CRC predisposition genes

in an unbiased manner, we analyzed WES data from 22 unexplained early-onset CRC patients and studied 95 familial CRC patients as a validation set. Cases with known hereditary CRC syndromes were excluded. Minor allele frequencies (MAFs) were estimated in 3,374 Finnish and 58,112 non-Finnish controls. In this series of 22 early-onset cases, we did not find any genes with recurrent loss-of-function (LoF) variants with MAF <0.1%. This observation, together with negative family history in 86% (19/22) of the unexplained young patients, suggests that the genetic background of these patients may be complex. Rare LoF variants in three genes - *ADAMTS4*, *CYTL1* and *SYNE1* - were shared between early-onset and familial CRC cases. Both *INTS5* and *ACSL5* harbored rare missense variants in two of the 22 patients, whereas *ARHGAP12*, *ATM*, *DONSON*, *MCTP2* and *ROS1* showed rare homozygous variants in single early-onset CRC cases. Further studies are needed to determine whether the identified variants are associated with CRC risk.

The third aim was to study the genetic basis of common, complex CRC. Each common variant has a small effect on disease risk, but their cumulative and population-level effects may be substantial. We conducted a genome-wide association study (GWAS) of CRC in 1,701 cases and 14,082 cancer-free controls from the Finnish population. Genotypes for a total of 9,068,015 common and low-frequency variants were imputed and analyzed, and most promising single-nucleotide polymorphisms (SNPs) were studied in additional 11,647 cases and 12,356 controls of European ancestry. The recently identified CRC risk SNP rs992157 on chromosome 2q35 was independently replicated ($p = 2.08 \cdot 10^{-4}$; odds ratio (OR), 1.14; 95% CI, 1.06-1.23), and it showed a genome-wide significant association in combined analysis ($p = 1.50 \cdot 10^{-9}$). Twelve additional loci (6p21.2, 8q23.3, 8q24.21, 10q22.3, 10q24.2, 11q13.4, 11q23.1, 14q22.2, 15q13.3, 18q21.1, 20p12.3 and 20q13.33) were associated with CRC in the Finnish population (false discovery rate <0.1), which replicates the associations of these loci with CRC and underscores similarities in the genetic architecture of CRC susceptibility between the Finnish population isolate and outbred populations.

Review of the literature

1 Cancer

Cancer (from Latin: *crab*) refers to a heterogeneous group of diseases characterized by excessive and uncoordinated cell growth and division (Willis, 1952). Cancer may arise from the linings of body surfaces and cavities (carcinoma, mesothelioma), mesenchymal tissue (sarcoma), blood-forming tissue (lymphoma, leukemia), neuroectodermal tissue or germ cells. Most cancers are thought to originate from a single cell that gives rise to successively expanding clones (Nowell, 1976). Malignant tumors can invade adjacent tissues and spread to other parts of the body via lymphatics, blood vessels or direct seeding. Most cancer deaths are caused by metastases (Mehlen et al., 2006). Benign tumors, by definition, remain anatomically localized but may become clinically significant if they exert mechanical pressure on surrounding tissues or secrete excessive amounts of hormones. Malignant cells are incompletely differentiated, which manifests as abnormal size and shape, nuclear atypia and disturbed spatial orientation (Kumar et al., 2014). Fundamental capabilities that most cancer cells possess, the hallmarks of cancer, include proliferative signaling, evading growth suppressors, avoiding immune destruction, enabling replicative immortality, deregulating cellular energetics, resisting cell death, inducing angiogenesis and activating invasion and metastasis (Hanahan et al., 2011). The survival and proliferation of cancer cells in the body depends on interactions with many other cell types such as immune cells, fibroblasts, endothelial cells, pericytes and smooth muscle cells (Hanahan et al., 2011).

Mutations in cellular DNA have a strong influence on the pathogenesis of cancer, and cancer may thus be considered a genetic disease. Cancer genes are mainly involved in cellular processes that regulate cell survival, cell fate and genome integrity (Vogelstein et al., 2013). In addition to protein-coding genes, non-coding RNA genes are relevant to cancer development (Esteller, 2011). To date, mutations in 567 human genes (approximately 3% of all human genes) have been causally implicated in cancer (http://cancer.sanger.ac.uk/census). Mutations that confer a selective growth advantage are called driver mutations, whereas those that are neutral or harmful to the cell are called passenger mutations. In the majority of common cancers, driver mutations are outnumbered by passenger mutations.

Age is a fundamental risk factor for cancer. Epidemiologic studies by Nordling in 1953 and Armitage and Doll in 1954 suggested that six or seven cellular events are required for cancer development (Nordling, 1953; Armitage et al., 1954). This conclusion was based on a mathematical model that explained the relationship between age and cancer-related mortality. Tomasetti et al., 2015b, found that mutation count data from CRCs and lung cancers were consistent with only three driver events. In some cancers, the number of driver events may be only one or two (Vogelstein et al., 2013).

Cancer risk is influenced by environmental and hereditary factors. For most cancer types, environmental and behavioral factors are more important than hereditary factors (Lichtenstein et al., 2000). Nearly 20% of all new cancer cases can be attributed to tobacco smoking, and another 20% to suboptimal levels of 13 other environmental or behavioral factors: alcohol consumption, processed and red meat consumption, fruit and vegetable intake, fiber intake, salt intake, obesity, physical activity, occupational exposures, infections, radiation from the sun, ionizing radiation, exogenous hormones and breastfeeding (Parkin et al., 2011). Approximately 5% of common cancers are associated with hereditary cancer syndromes, whereas the contribution of common genetic variants is less clearly defined (Fletcher et al., 2010).

1.1 Oncogenes

Oncogenes encode proteins that are capable of promoting tumorigenesis. They are generally dominant at the cellular level - i.e., monoallelic gain-of-function contributes to tumorigenesis. Proto-oncogenes, which are normal counterparts of oncogenes, can be converted into oncogenes by mutation or overexpression. Oncogenes tend to be mutated at specific hotspots, but gene amplifications, translocations and gain-of-function deletions are also observed (Vogelstein et al., 2013; Zehir et al., 2017). Oncogenes include growth factors, protein kinases, intracellular signal transducers and transcription factors (Croce, 2008). For example, *KRAS* encodes an intracellular signaling protein that controls the Ras-mitogen activated protein kinase -pathway (Pylayeva-Gupta et al., 2011). *KRAS* is the most commonly mutated oncogene in human cancer (Zehir et al., 2017). *KRAS* is frequently mutated in adenocarcinomas of the pancreas, lung and colorectum (http://cancerhotspots.org/). Germline gain-of-function mutations in oncogenes are rare because of their embryonic lethality. An exception is the *RET* oncogene, which causes hereditary multiple endocrine neoplasia (MEN) type 2 when mutated (Mulligan et al., 1993; Hofstra et al., 1994).

1.2 Tumor suppressor genes

Tumor suppressor genes (TSGs) encode proteins that are capable of preventing tumorigenesis. They are generally recessive at the cellular level - i.e., biallelic LoF contributes to tumorigenesis. Mutations in tumor suppressor genes are characteristically spread across exons and exon-intron boundaries, and the proportion of protein-truncating nonsense, splice site or frameshift mutations is high (Vogelstein et al., 2013). Truncated genes may be functionally impaired or underexpressed due to nonsense-mediated RNA decay (Conti et al., 2005). TSGs can be divided into gatekeepers that directly limit tumor initiation (e.g., *RB1*, *VHL*, *NF1* and *APC*) and caretakers that maintain genome integrity (e.g., *MLH1*, *BRCA1*, *BRCA2* and *ATM*) (Kinzler et al., 1997). Germline mutations in TSGs cause autosomal dominant cancer susceptibility syndromes (Vogelstein et al., 2004). Patients who carry germline mutations in TSGs are at increased risk of cancer because only one additional somatic mutation is required for biallelic LoF. This two-hit hypothesis was originally proposed by Alfred Knudson who analyzed a series of 48 cases of retinoblastoma (Knudson, 1971). Fifteen years later, it was found that genetic susceptibility to early-onset retinoblastoma (as well as osteosarcoma) is caused by germline mutations in *RB1*, which is a negative regulator of the cell cycle (Friend et al., 1986; Giacinti et al., 2006). The most common mechanism for the inactivation of the wild-type allele of TSGs is loss of heterozygosity (LOH). LOH may be due to chromosomal deletion, mitotic recombination or uniparental disomy; the latter two mechanisms preserve the copy number of the gene (Ryland et al., 2015). Some TSGs are haploinsufficient; in this case, the loss of one gene copy has a tumorigenic effect (Santarosa et al., 2004). Haploinsufficiency is to be distinguished from a dominant-negative effect, in which mutant proteins interfere with normal copies (Chenevix-Trench et al., 2002). Epigenetic silencing of TSGs can occur by biallelic promoter hypermethylation, which is known to affect *MLH1* in CRC, *VHL* in renal cell cancer and *CDKN2A* in multiple cancer types (Kane et al., 1997; Merlo et al., 1995; Herman et al., 1994).

1.3 Genomic instability

Somatic mutations are generated by DNA replication, chemical mutagens, radiation, viruses and retrotransposons. Because of tissue self-renewal and environmental exposures, somatic mutations accumulate in normal tissues during aging (Martincorena et al., 2015). Genomic instability, which is present in most cancers, refers to an elevated rate of point mutations or chromosomal aberrations (Negrini et al., 2010). It may be caused by impaired chromosome segregation, homologous recombination, mismatch repair (MMR), base- or nucleotide-excision repair or polymerase proofreading (Thompson et al., 2010; Harris, 2013). Under normal circumstances, DNA damage signaling prevents genetically damaged cells from proliferating (Jeggo et al., 2016). The most common defect in DNA damage signaling is inactivation of the p53 transcription factor. The corresponding gene, *TP53*, is the most frequently mutated gene in human cancer (Zehir et al., 2017). The p53 pathway is normally activated in response to DNA damage and oncogenic stress (Reinhardt et al., 2012). When activated, p53 may trigger cell cycle arrest, cellular senescence or apoptosis (Reinhardt et al., 2012). Germline mutations in *TP53* underlie the Li-Fraumeni syndrome, which is associated with increased risks of breast cancer, sarcoma and multiple other tumor types (Malkin, 2011).

2 Genetic predisposition to cancer

Clinical signs of hereditary cancer syndromes include familial aggregation, early age of onset, multiple primary cancers and non-malignant manifestations such as café-au-lait spots in neurofibromatosis type 1 (Nagy et al., 2004). ”Cancer families” were first documented in the early 1900s (Peutz, 1921; Warthin, 1931; Helwig, 1946; Jeghers et al., 1949). The corresponding genetic defects, however, remained elusive until the 1990s when genetic linkage and positional cloning studies became feasible. To date, mutations in approximately 100 genes have been implicated in inherited susceptibility to human cancer (`cancer.sanger.ac.uk/`

census). Genetic studies of rare hereditary syndromes have contributed to our understanding of the pathogenesis of common cancers that lack a strong heritable component (Kinzler et al., 1996). On the other hand, GWASs have identified more than 450 common SNPs that are associated with low-penetrance cancer predisposition (Sud et al., 2017). In recent years, next-generation sequencing (NGS) has attracted interest due to its potential for identifying disease-causing variants that are too rare to be tagged by common SNPs but too weakly penetrant to be detected by linkage analysis (Zuk et al., 2014).

2.1 Familial aggregation and heritability

Most cancer types show statistically significant familial aggregation (Goldgar et al., 1994; Hemminki et al., 1998b). The familial relative risk (RR) for first-degree relatives of affected individuals is typically two- to threefold, but some cancers such as testicular cancer, thyroid cancer and chronic lymphocytic leukemia display even stronger familial aggregation (Goldgar et al., 1994; Hemminki et al., 1998b; Goldin et al., 2004). Familial clusters are caused by both environmental and hereditary factors. To study the relative contributions of these influences, it is useful to define heritability as the proportion of phenotypic variance that can be ascribed to hereditary factors. For categorical traits, phenotypic variance may refer to indirectly observed variance on a normally distributed liability scale (Visscher et al., 2008). Heritabilities of common cancers have been estimated in twin studies and GWASs, and statistically significant hereditary effects have been found for prostate cancer, breast cancer, CRC, testicular cancer, kidney cancer, melanoma and non-melanoma skin cancer (Sampson et al., 2015; Mucci et al., 2016; Graff et al., 2017).

2.2 Linkage and association studies

Genetic linkage studies

Genetic linkage studies aim to identify polymorphic markers that are physically linked to a disease susceptibility locus (Dawn Teare et al., 2005). A set of markers is selected to capture variation across the genome, and a statistical test is carried out for each marker. No prior knowledge on candidate loci is required.

Under the null hypothesis of no linkage, the recombination fraction (θ) between a marker locus and a causal locus is $\frac{1}{2}$. In the case of parametric linkage analysis, a likelihood ratio statistic is constructed to compare null likelihood ($\theta = \frac{1}{2}$) with maximized likelihood ($0 \leq \theta \leq \frac{1}{2}$) (Haldane et al., 1947). The 10-based logarithm of this statistic (the logarithm of the odds (LOD) score) is compared to a prespecified significance threshold. Under certain assumptions, a LOD score of 3 corresponds to a 90% posterior probability of linkage between the marker and disease susceptibility locus (Sham, 1998). Non-parametric linkage analysis is a model-free statistical approach that may be used to avoid restrictive assumptions on penetrance, mode of inheritance or other parameters (Balding et al., 2007). Data from multiple families can be combined to increase statistical power, and sufficiently narrow linkage regions can be analyzed for mutations. An illustrative example of the application of linkage analysis in cancer research is the mapping of a CRC susceptibility locus to chromosome 2p (Peltomäki et al., 1993). In this study, a total of 345 microsatellite markers were analyzed in two large kindreds (family C from North America and family J from New Zealand), which resulted in the identification of the marker D2S123 with estimated $\theta = 0$ in both families. The pairwise LOD scores in families C and J were 6.39 and 1.45, respectively, providing strong evidence of linkage. Both families were subsequently found to harbor pathogenic germline mutations in *MSH2* (Leach et al., 1993).

Genome-wide association studies

Even before GWASs became feasible, it was proposed that genetic association studies would detect small genetic effects with greater statistical power than genetic linkage studies (Risch et al., 1996). In 2005, the International HapMap Project produced a genomic map of linkage disequilibrium (LD), which showed that the majority of common variation could be tagged with a relatively small number of common SNPs (International HapMap Consortium, 2005). Data on genome-wide LD patterns informed the development of SNP arrays, which, together with large sample series, led to a surge in the number of published GWASs beginning in 2007 (Visscher et al., 2017). Similar to genetic linkage studies, GWASs require no prior knowledge of the location or biological nature of disease susceptibility loci and may thus reveal previously unsuspected biological mechanisms.

Most GWASs of cancer susceptibility are case-control studies in which hundreds of thousands to millions of SNPs are genotyped across the genome. To avoid bias, cases and controls should be sampled from the same reference population, and DNA samples from cases and controls should be genotyped and processed with similar biochemical and bioinformatic methods (Wang et al., 2005). Cases may be selected for high-risk features such as early-onset or familial cancer. Similarly, controls may be selected for low-risk features such as being cancer-free at old age. Due to the relative rarity of any particular cancer type, statistical power can be increased by including cases with premalignant lesions such as colorectal adenomas in studies of CRC or Barrett's esophagus in studies of esophageal cancer (Tomlinson et al., 2007; Gharahkhani et al., 2016). Pleiotropic effects on multiple cancer types have been observed for some SNPs (e.g., rs6983267 near *MYC* at 8q24 and rs2736100 near *TERT* at 5p15.33), which supports the idea of investigating composite phenotypes of multiple genetically related cancers (Tomlinson et al., 2007; Thomas et al., 2008; Rafnar et al., 2009). Because of economic constraints, primary GWASs may be followed by one or more replication stages in which smaller numbers of promising SNPs are investigated in additional case-control series. To increase statistical power, the primary test for disease association can be based on combined analysis of discovery and replication sets (Skol et al., 2006). Because of local correlations between SNPs (LD), genome-wide testing of all common variants is roughly equivalent to 10^6 independent statistical tests (Pe'er et al., 2008). At a type 1 error rate of 5%, this corresponds to a Bonferroni-corrected significance threshold of $\frac{0.05}{10^6} = 5 \cdot 10^{-8}$. Strategies for functional characterization of GWAS-derived risk loci include fine-mapping, resequencing, regulatory annotation, expression quantitative trait locus analysis and experimental models (Freedman et al., 2011).

Because the effects of common SNPs on cancer risk are small, meta-analysis of multiple primary GWASs is a useful strategy for increasing statistical power and eliminating study-specific false-positive results (Evangelou et al., 2013). Genotype imputation helps to harmonize datasets and to recover genotypes that are missing in a subset of studies (McCarthy et al., 2016). Meta-analyses may be based on fixed-effect or random-effects models (Riley et al., 2011). In fixed-effect meta-analysis, the genetic effect is assumed to be the same in each study. In many practical settings, however, the genetic effect is somewhat variable because of differences in study designs, methods and populations. Random-effects meta-analysis allows the genetic effect to vary from study to study, and it can be used to estimate the mean genetic effect in a population of studies. In the absence of study heterogeneity, fixed-effect and random-effects models yield numerically identical results. Meta-analysis of GWASs can be performed with summary statistics, which improves computational efficiency and protects the privacy of study participants (Pasaniuc et al., 2017).

Common SNPs identified in GWASs have small effects on cancer risk, but their cumulative and population-level effects may be substantial. The combined effects of multiple risk alleles can be quantified with the use of polygenic risk scores (PRSs), which, under some assumptions, model the relationship between genotype combinations (2^n for n variants) and disease risk. As compared to the population median, RRs for individuals in the highest percentile of PRSs have been estimated at 4.7, 3.2 and 2.9 for prostate, breast and colorectal cancers, respectively (Eeles et al., 2013; Michailidou et al., 2013; Frampton et al., 2016). The predictive accuracy of PRSs can be improved by adding further risk SNPs and by including information on family history (Chatterjee et al., 2013). Also, causal SNPs (which are often unknown) are likely to be more informative than tag SNPs that are used in GWASs. An area of ongoing investigation is whether exponential (log-additive) or linear (additive) models should be used to model cumulative effects; the exponential model implies higher risks for individuals with large numbers of risk alleles (Kraft, 2017).

2.3 Next-generation sequencing

NGS is a high-throughput technology that has dramatically reduced the cost and run time for generating large-scale genomic data. To put this in perspective, it has been estimated that Sanger sequencing of a human genome on one machine takes approximately 60 years, whereas NGS-based whole-genome sequencing (WGS) can be performed in one or two days (Bennett et al., 2005; Miller et al., 2015). NGS typically begins with library preparation, in which DNA is purified and randomly fragmented (Metzker, 2010). Polymerase chain reaction (PCR)-based amplification can increase signal strength during sequencing but also introduces polymerase errors and amplification bias (e.g., underrepresentation of regions with high GC-content; Aird et al., 2011). Unless the whole genome is sequenced, targeted regions are captured with probe hybridization to reduce sequencing costs. Next, DNA fragments are typically immobilized on solid surfaces or beads. The prepared DNA fragments can be sequenced with a variety of methods. In sequencing-by-synthesis, fluorescent or otherwise labeled nucleotides produce base-specific signals when they are incorporated into a complementary DNA strand. The efficiency of NGS is explained by the massive parallelization of the sequencing process. The sequencing reads are mapped into a reference genome or assembled *de novo* to produce variant calls.

Progress in NGS has revolutionized many areas of cancer research and clinical oncology. The Cancer Genome Atlas project has produced WES data from thousands of tumors and matched normal tissues, which has enabled the discovery of significantly mutated cancer genes across more than 20 tumor types

(Lawrence et al., 2014). These results indicate that while some cancer genes are relatively tissue-specific, others are frequently mutated in a variety of tumor types. In addition, there is substantial genomic diversity among tumors within the same histopathologic category. Because exomes contain only part of the information that may be gained from cancer genomes, the Pan-Cancer Analysis of Whole Genomes consortium has collected and harmonized WGS data from more than 2,600 tumors and is beginning to reveal genomic patterns of somatic structural variants, copy number alterations and non-coding point mutations (https://www.biorxiv.org/content/biorxiv/early/2017/07/12/162784). In clinical oncology, the MSK-IMPACT panel was used to perform somatic mutation profiling in a prospective series of more than 10,000 patients with metastatic cancer (Zehir et al., 2017). MSK-IMPACT revealed at least one clinically actionable mutation in 38% of the patients. On the other hand, NGS has revealed a number of novel tumor susceptibility syndromes, which not only improves genetic diagnostics but also advances our understanding of general disease mechanisms (Jamshidi et al., 2015). Large-scale projects are underway to realize the potential of NGS in complex disease genetics such as multifactorial cancer susceptibility (Goldfeder et al., 2017).

2.4 Genetic testing for cancer risk

The purpose of genetic testing for cancer risk is to identify individuals who are genetically predisposed to cancer and may benefit from clinical intervention. Genetic counseling is offered before and after genetic testing to obtain informed consent, to interpret and explain results, and to give recommendations on clinical management. Assessment of genetic risk is a common clinical challenge in the management of CRC and breast cancer (https://www.nccn.org/). Family history is used to estimate mutation probability and to prioritize family members for genetic testing. An accurate pedigree should capture all cancer cases and unaffected relatives in three or four generations, including information on age at cancer diagnosis and benign conditions associated with hereditary cancer syndromes (Rimoin et al., 2013). Although hereditary syndromes can be suspected on the basis of family history and tumor features, the sensitivity of guideline-directed genetic testing is suboptimal (Mandelker et al., 2017).

Sanger sequencing and multiplex ligation-dependent probe amplification (MLPA) are cost-effective methods for testing one or a small number of genes, whereas NGS-based methods such as WES, WGS and targeted gene panels can screen large numbers of genes in parallel. Diagnostic NGS may be useful when a genetic disorder is suspected but clinical features do not suggest a specific syndrome (Robson et al., 2015). A challenge inherent in multiplex testing is the elevated frequency of variants of uncertain significance (VUSs) and incidental findings. VUSs can be reanalyzed periodically and reclassified when sufficient evidence has accumulated (Richards et al., 2015). According to the recommendations of the American College of Medical Genetics and Genomics, incidental findings in all clinically actionable genes should be reported, regardless of the clinical context (Kalia et al., 2017). It has been suggested, however, that especially cancer patients may have limited time and energy to manage information on incidental findings, and therefore an opt-out policy may be preferred (Parsons et al., 2014). If genetic testing is restricted to firmly established cancer susceptibility genes, incidental findings are also likely to be related to cancer susceptibility. NGS can also be used for somatic tumor profiling, which often yields secondary and incidental information on germline predisposition. This possibility needs to be discussed with cancer patients prior to tumor profiling (Robson et al., 2015).

3 Colorectal cancer

3.1 Epidemiologic and clinical aspects

Worldwide, CRC accounts for approximately 10% of the 14 million new cancers that are diagnosed each year and a similar proportion of cancer deaths. In 2012, the global age-standardized incidence rate of CRC was 21 per 100,000 person-years for males and 14 per 100,000 person-years for females with wide geographic variation (Ferlay et al., 2015). For example, the incidence of CRC was highest in Australia (44.8 per 100,000 person-years for males and 32.2 per 100,000 person-years for females) and lowest in Western Africa (4.5 per 100,000 person-years for males and 3.8 per 100,000 person-years for females). Age is the strongest demographic risk factor for CRC. An estimated eleven percent of CRCs are diagnosed before age 50 years and 3.2% before age 40 years. In developed countries, the cumulative risk of CRC by age 75 years was 4.3% for males and 2.7% for females.

Environmental and behavioral risk factors for CRC include obesity, physical inactivity, high consumption of red and processed meat, low fiber intake, low intake of fruits and vegetables, high alcohol consumption and smoking (Bouvard et al., 2015; Larsson et al., 2007; Lee et al., 2012; Aune et al., 2011a; Botteri et al., 2008; Aune et al., 2011b; Cho et al., 2004; Koushik et al., 2007).

There are also medical risk factors for CRC. Inflammatory bowel disease (IBD) is associated with a twofold overall risk of CRC, but personal risk depends on the presence of dysplasia, duration of disease, extent of inflammation and anatomical complications (Jess et al., 2012; Lutgens et al., 2013; Beaugerie et al., 2015). Data from randomized clinical trials (RCTs) suggest that non-steroidal anti-inflammatory drugs and postmenopausal hormone therapy are associated with decreased risk of CRC, whereas diabetes mellitus was associated with increased risk of CRC in a meta-analysis of observational studies including both type 1 and type 2 diabetics (Bibbins-Domingo et al., 2016; Simon et al., 2012; Deng et al., 2012).

Approximately 10% of CRC cases have at least one first-degree relative with CRC (Salovaara et al., 2000). In a meta-analysis of 26 case-control and cohort studies, first-degree relatives of CRC patients had a 2.25-fold RR of CRC (95% CI, 2.00-2.53) as compared with those with negative family history (Johns et al., 2001). The risk was higher for those with multiple or early-onset CRC cases in the family. A significant part of the familial aggregation of CRC is related to hereditary factors; in a Nordic twin study, the heritability of CRC was estimated at 40% (95% CI, 33%-48%; Graff et al., 2017).

There is strong evidence that screening for CRC in average-risk adults aged 50 to 75 years reduces disease-specific mortality, and that the benefits outweigh the harms, which are mainly related to endoscopic complications (Lin et al., 2016). Screening decisions in those younger than 50 years or older than 75 years should be made on an individual basis. Persons with affected first-degree relatives may begin screening at age 40 years or 10 years earlier than the earliest CRC case in the family (Rex et al., 2017). Among a number of possible screening methods, two are currently supported by RCTs: flexible sigmoidoscopy and guaiac-based fecal occult blood test (gFOBT). In meta-analyses of RCTs (four studies on flexible sigmoidoscopy and five studies on gFOBT), flexible sigmoidoscopy reduced CRC-specific mortality with an incidence rate ratio (IRR) of 0.73 (95% CI, 0.66-0.82), and gFOBT reduced CRC-specific mortality with a RR of 0.91 at 19.5 years (95% CI, 0.84-0.98) and 0.78 at 30 years (95% CI, 0.65-0.93) (Lin et al., 2016). Other screening methods available for clinical use are based on either direct visualization (colonoscopy and computed tomographic colonography) or stool analysis (fecal immunochemical test (FIT) and stool DNA). In Finland, population-level screening for CRC was evaluated in a randomized health services study during 2004-2016. The study included individuals aged 60-69 and compared biennial gFOBT with no organized screening. At a median follow-up of 4.5 years (based on data from years 2004-2012; Pitkäniemi et al., 2015), the incidence of CRC had increased in the screening arm (IRR, 1.11; 95% CI, 1.01-1.23), but there was no significant difference in CRC-specific mortality between the screening and control arms (IRR, 1.04; 95% CI, 0.84-1.28). Longer follow-up is required to draw more definite conclusions regarding the effect on CRC-specific mortality. A new screening program, based on biennial FIT, is planned to begin in Finland in 2019 and is expected to expand from volunteering municipalities into a nationwide program.

Five-year relative survival rates for patients with CRC have improved over the last decades and already exceed 60% in developed countries (Miller et al., 2016). In the United States in 2007-2013, five-year relative survival rates were 90% for localized disease, 71% for regionally metastatic disease and 14% for distant metastatic disease, underscoring the value of early diagnosis. In the 8th edition of the Tumor-Node-Metastasis (TNM) classification, CRC is staged into eleven anatomic-prognostic groups (0, I, IIA-C, IIIA-C and IVA-C; Amin et al., 2016). The older Dukes' (A-D) and modified Astler-Coller (A, B1-3, C1-3, D) systems are not recommended for clinical use anymore. In addition to advanced TNM stage, unfavorable tumor characteristics include *BRAF* p.Val600Glu mutation (Modest et al., 2016), lymphovascular invasion (Hogan et al., 2015), perineural invasion (Liebig et al., 2009), poor differentiation (Chapuis et al., 1985), signet ring cell histology (Nissan et al., 1999) and clinical presentation with obstruction or perforation (Chen et al., 2000).

The primary treatment for most patients with resectable CRC is surgery with curative intent. Depending on the risk of recurrent disease and surgical criteria, surgery may be combined with preoperative, perioperative or postoperative oncological treatment (Van Cutsem et al., 2016). Patients with unresectable disease who are candidates for systemic therapy may be treated with conventional chemotherapy (fluoropyrimidines, oxaliplatin, irinotecan), antiangiogenic agents (bevacizumab, ramucirumab, aflibercept), epidermal growth factor receptor antibodies (cetuximab, panitumumab) or the multi-kinase inhibitor regorafenib (Van Cutsem et al., 2016). Immunotherapy is being actively investigated in MSI CRC (Le et al., 2017), whereas anti-*HER2* therapy has shown promise in *HER2*-positive metastatic CRC (Sartore-Bianchi et al., 2016).

3.2 Molecular pathogenesis

The molecular pathogenesis of CRC is a multistage process that may take years or decades. The adenoma-carcinoma sequence describes the stepwise accumulation of genetic events that lead from normal colorectal epithelium to invasive adenocarcinoma (Fearon et al., 1990; Luebeck et al., 2002). Most driver events are somatic, but some can be inherited. Early studies showed allelic losses on chromosomes 5q (containing *APC*), 8p, 17p (containing *TP53*) and 18q (containing a cluster of *SMAD* genes and *DCC*), suggesting the presence of TSGs in these chromosomal regions (Vogelstein et al., 1989). Loss of *APC* is very common even in smallest adenomas and initiates tumorigenesis by activating the Wnt-β-catenin signaling pathway (Powell et al., 1992). *APC*-wild-type CRCs (20-30%; Muzny et al., 2012) may harbor *CTNNB1* mutations or R-spondin fusion genes (*RSPO2* or *RSPO3*), which provide alternative routes for activating Wnt-β-catenin signaling (Morin et al., 1997; Seshagiri et al., 2012). *KRAS* mutations are common in larger adenomas (>1 cm) and invasive adenocarcinomas (Forrester et al., 1987; Boland et al., 1995). *KRAS* mutations, which are present in 40% of CRCs, are mutually exclusive with *BRAF* p.Val600Glu, which is observed in 10% of CRCs (Muzny et al., 2012). *BRAF* encodes a serine-threonine kinase downstream of *KRAS* in the mitogen-activated protein kinase pathway (Vogelstein et al., 2004). It remains poorly understood why *BRAF*-mutant CRCs are clinically more aggressive than *KRAS*-mutant CRCs. *BRAF* mutations outside codon p.Val600 account for one fifth of *BRAF* mutations in CRC and seem to be associated with more favorable prognosis than *BRAF* p.Val600Glu (`http://www.cancerhotspots.org/`; Jones et al., 2017). Loss of *TP53* is typically a late event in colorectal carcinogenesis and correlates with transition from adenoma to carcinoma (Ohue et al., 1994; Boland et al., 1995). The phosphoinositide-3-kinase and transforming growth factor (TGF)-β signaling pathways are also frequently dysregulated (Muzny et al., 2012).

Ten to 15 percent of CRCs display MSI (Aaltonen et al., 1998), which is caused by inherited or somatic defects in DNA MMR genes (*MLH1*, *MSH2*, *MSH6* and *PMS2*). MSI CRCs are usually diploid or near-diploid but display large numbers of unrepaired replication errors, which are frequent at mono-, di- and trinucleotide repeat tracts (Thibodeau et al., 1998). Underexpression of MMR genes has been observed in aberrant crypt foci that may be early precursors to MSI CRC (Leggett et al., 2010; Kloor et al., 2012). Inactivation of the TGF-β signaling pathway by frameshift mutations in *TGFBR2* occurs in as many as 90% of MSI CRCs (Kondelin et al., 2017). Mutations in *KRAS* and *TP53*, however, are less frequent in MSI CRC than in MSS CRC (Sinicrope et al., 2013; Lin et al., 2015). Although global hypomethylation is a common characteristic of CRC, CpG islands (including the promoter region of *MLH1*) are highly methylated in a subset of CRCs that express the CpG island methylator phenotype (Goelz et al., 1985). The CpG island methylator phenotype is often associated with *BRAF* p.Val600Glu, tumor MSI and serrated histology (Leggett et al., 2010). Somatic *BRAF* p.Val600Glu mutation is rare in patients with germline mutations is MMR genes (Domingo et al., 2004), which underlines biological differences between sporadic and LS-related MSI CRC. Clinicopathologic features associated with MSI CRC include proximal location, mucinous histology, poor differentiation, immune cell infiltration and lower risk of distant metastases (Kim et al., 1994; Kakar et al., 2003; Malesci et al., 2007). MSI CRC is common in female and elderly patients (Kakar et al., 2003).

In recent years, NGS-based tumor profiling has advanced our understanding of the genetic landscape of CRC. In exome-wide studies, the median numbers of nonsynonymous mutations in MSS and MSI CRCs have been estimated at 66 (interquartile range (IQR), 47-92) and 777 (IQR, 494-1,326), respectively (Wood et al., 2007; Muzny et al., 2012). Most of the observed mutations appear to be passenger events. With the use of recently developed statistical methods, new driver genes have been identified and linked to incompletely understood biological processes such as chromatin remodeling (*ARID1A*), proteolysis (*FBXW7*) and RNA processing (*PCBP1*) (Muzny et al., 2012; Lawrence et al., 2014; Kandoth et al., 2013; Domingo et al., 2016). In particular, somatic mutations in the proofreading domain of *POLE* have been found in ultramutated CRCs that represent approximately 1% of all CRCs (Muzny et al., 2012). On the other hand, NGS has enabled large-scale analyses of somatic mutation processes. Mutation signatures associated with CRC include signature 1B (associated with aging), signature 6 (associated with MSI), signature 10 (associated with *POLE* mutations) and signature 17 (associated with unknown etiology) (Alexandrov et al., 2013; Katainen et al., 2015). Moreover, gene expression studies have classified CRCs into four consensus molecular subtypes that cannot be explained by any single genetic defect alone (*CMS1*, "MSI immune"; *CMS2*, "canonical"; *CMS3*, "metabolic"; and *CMS4*, "mesenchymal"; Guinney et al., 2015).

3.3 Hereditary colorectal cancer syndromes

Cancers of the GI tract account for approximately 29% of all incident cancers and 37% of all cancer deaths globally (Ferlay et al., 2015). Environmental exposures are considered the principal etiologic factors for most GI cancers. The contribution of genetic factors, however, appears to be important especially for cancers of the colon and rectum (as compared to cancers of the upper GI tract). Hereditary CRC syndromes can be broadly divided into hereditary nonpolyposis colorectal cancer (HNPCC; now termed LS) and GI polyposis syndromes. These syndromes are reviewed in this chapter and summarized in Table 3.1 and Figure 3.1.

3.3.1 Lynch syndrome

Genetic basis. LS is the most common hereditary CRC syndrome, which accounts for approximately 3% of new CRC cases (Moreira et al., 2012). Its estimated prevalence in North American and Australian populations is 1:279 (95% CI, 192-403 Win et al., 2017), whereas the prevalence may be higher in certain founder populations (Boland et al., 2018). The exact prevalence in Finland is unknown (Chapelle, 2005). In 1993, tumor MSI was detected in families with autosomal dominant CRC susceptibility, and underlying germline mutations in DNA MMR genes were identified soon thereafter (Aaltonen et al., 1993; Peltomäki et al., 1993; Fishel et al., 1993; Leach et al., 1993; Bronner et al., 1994; Papadopoulos et al., 1994; Nicolaides et al., 1994; Miyaki et al., 1997). MMR proteins participate in DNA repair by forming heterodimers that recognize base-pair mismatches (Kunkel et al., 2005). An uncommon cause of LS is deletion of the 3' end of *EPCAM* (chromosome 2p21), which leads to constitutional hypermethylation of the promoter region of *MSH2 in cis* (Ligtenberg et al., 2009; Kovacs et al., 2009).

Clinical manifestations. The penetrance of LS depends on the mutated gene and sex. In a prospective study of 3,119 patients who were followed for 24,475 person-years, the cumulative risk of CRC by age 75 years was 46%, 43% and 15% for germline mutations in *MLH1* (n=1,473), *MSH2* (n=1,060) and *MSH6* (n=462), respectively (Møller et al., 2017). Patients with mutations in *PMS2* (n=124) remained CRC-free during the 524 person-years of follow-up, reflecting the lower penetrance of this gene defect (Møller et al., 2017). Women with LS appear to be at similar risk of CRC and endometrial cancer. Other LS-associated cancers include gastric, small bowel, bile duct, renal, urothelial, ovarian and brain cancers (Watson et al., 2005). Muir-Torre syndrome refers to the co-occurrence of skin tumors (keratoacanthomas and sebaceous gland tumors) and internal malignancies and is often a subtype of LS (Ponti et al., 2005). The number of colorectal adenomas in patients with LS is somewhat higher than in the general population but typically remains below 10 (Kalady et al., 2015). Biallelic germline mutations in MMR genes (often in *PMS2*) cause a syndrome of constitutional MMR deficiency, which is associated with CRC, small bowel cancer, brain tumors, lymphoma and leukemia that often manifest in childhood (Vasen et al., 2014).

Diagnosis and management. The revised Bethesda Guidelines and Amsterdam II criteria can be used to select patients for MSI testing (Vasen et al., 1999; Umar et al., 2004). Because of the imperfect sensitivities of these clinical criteria (87.8% and 27.2%, respectively; Moreira et al., 2012), clinical practice has shifted towards universal MSI screening among newly diagnosed CRC patients (Heald et al., 2013). Somatic *BRAF* p.Val600Glu mutation excludes LS with high confidence, which facilitates universal screening (Toon et al., 2013). MSI testing can be performed with MMR protein immunohistochemistry (IHC), direct analysis of the Bethesda markers (BAT26, BAT25, D5S346, D2S123 and D17S250) or NGS-based classifiers (Suraweera et al., 2002; Hampel et al., 2005; Hause et al., 2016). IHC is inexpensive and provides information on which MMR gene is likely to be mutated. Screening for germline mutations in MMR genes is recommended for all patients with MSI CRC, and the identification of a pathogenic germline variant in *MLH1*, *MSH2*, *MSH6*, *PMS2* or *EPCAM* (`https://www.insight-group.org/variants/databases/`) establishes the diagnosis of LS. Surveillance colonoscopy is recommended every one or two years beginning at age 20-25 years (or even earlier, depending on family history) to reduce CRC-specific and overall mortality (Giardiello et al., 2014; Syngal et al., 2015; Järvinen et al., 2000). Aspirin (600 mg per day) appeared to reduce the incidence of LS-related CRC in a RCT, but further confirmatory studies are required (Burn et al., 2011). The prognosis of LS-associated CRC is favorable with a 10-year overall survival rate of 91% (Møller et al., 2017). Women with LS should be offered surveillance for endometrial and ovarian cancers, as well as prophylactic hysterectomy and bilateral salpingo-oophorectomy after child-bearing age (Giardiello et al., 2014; Syngal et al., 2015).

3.3.2 Gastrointestinal polyposis syndromes

Familial adenomatous polyposis

Genetic basis. FAP is the second most common hereditary CRC syndrome (Gardner, 1951). Its incidence is approximately 1 per 10,000 live births, and it accounts for <1% of newly diagnosed CRCs (Björk et al., 1999; Bülow, 2003). An important first clue to the genetic etiology of FAP was a 42-year-old man with colon cancer, colorectal polyposis, congenital malformations and an interstitial deletion between 5q13-q22 (Herrera et al., 1986). In 1987, the gene underlying FAP was mapped near 5q21-q22 (Bodmer et al., 1987). In 1991, the simultaneous efforts of two research groups revealed germline and somatic mutations in *APC* (Kinzler et al., 1991; Nishisho et al., 1991; Groden et al., 1991; Joslyn et al., 1991). As many as 25% of patients with FAP present with *de novo* mutations and a negative family history (Bisgaard et al., 1994).

Clinical manifestations. Patients with classical FAP begin to develop adenomas in their adolescent years, and the number of polyps is likely to increase to more than 100 by adulthood. Untreated patients with classical FAP are at 90% cumulative risk of developing CRC by age 50 years, but this risk can be greatly reduced with prophylactic surgery and endoscopic surveillance (Bussey, 1975). Patients with smaller numbers of polyps (10-100) and later age at presentation (often $\geq$ 50 years) are classified as having attenuated FAP (Knudsen et al., 2003). Attenuated FAP may be caused by mutations in the 5' end (proximal to codon 158) or 3' end (distal to codon 1,596) of *APC* (Brensinger et al., 1998; Heppner Goss et al., 2002). Extraintestinal manifestations also depend on the site of mutation (Galiatsatos et al., 2006). In addition to CRC, FAP is associated with medulloblastoma, hepatoblastoma, papillary thyroid cancer, duodenal cancer, ampullary cancer, pancreatic cancer and gastric cancer (Galiatsatos et al., 2006). An important extraintestinal manifestation of FAP is desmoid tumor (mesenteric fibromatosis), which causes a high degree of morbidity and mortality (Slowik et al., 2015).

Diagnosis and management. Genetic testing for *APC* mutations is recommended in individuals with a personal history of at least 10-20 colorectal adenomas, typical extraintestinal manifestations (desmoid tumor, hepatoblastoma, papillary thyroid cancer or congenital hypertrophy of the retinal pigment epithelium) or a known mutation in the family (`https://www.nccn.org/`). Annual surveillance with flexible sigmoidoscopy or colonoscopy is usually begun at puberty (Syngal et al., 2015). The primary prophylactic treatment for classical FAP is proctocolectomy or colectomy in the second or third decade of life. Rectal endoscopy is recommended at least once every year after rectum-preserving surgery. The approach to surveillance and prophylactic surgery in attenuated FAP is less aggressive; annual colonoscopic surveillance can be initiated at age 25 years, and all patients may not require prophylactic colectomy.

MUTYH-associated polyposis

MUTYH-associated polyposis (MAP) is an autosomal recessive syndrome characterized by multiple colorectal adenomas (typically 3-100) and high-penetrance CRC predisposition (Sieber et al., 2003). MAP-related colorectal tumors show a high frequency of somatic G>T transversions due to impaired base-excision repair (Al-Tassan et al., 2002). The two most frequent mutations in *MUTYH* are p.Tyr176Cys (rs34612342) and p.Gly393Asp (rs36053993), which have a combined allele frequency of 1% in European populations (`http://gnomad.broadinstitute.org`). In biallelic mutation carriers, the cumulative risk of CRC by age 70 years has been estimated at 75.4% for males and 71.7% for females (Win et al., 2014). The risk of CRC in monoallelic *MUTYH* mutation carriers may be moderately elevated (Lubbe et al., 2009; Win et al., 2014). Extracolonic manifestations of MAP include gastric and duodenal polyps and duodenal cancer (Nielsen et al., 2005; Vogt et al., 2009). The clinical management of MAP is roughly similar to attenuated FAP (Syngal et al., 2015).

Juvenile polyposis

JP is an autosomal dominant syndrome that predisposes to CRC, juvenile polyps and congenital anomalies (Calva et al., 2008). Data on its incidence are scarce, but the incidence is thought to be approximately 1:100,000-1:160,000 (Syngal et al., 2015). In 1998, Howe et al. mapped a gene for JP to chromosome 18q21.1, which led to the identification of a pathogenic frameshift variant in *SMAD4* (Howe et al., 1998a; Howe et al., 1998b). In 2001, linkage analysis identified another JP locus near 10q22-q23, and nonsense mutations were detected in *BMPR1A*, which, similar to *SMAD4*, is a component of the TGF-β signaling pathway (Howe et al., 2001). Nonetheless, as many as 60% of JP cases remain genetically unexplained (Howe et al., 2004). The estimated risk of CRC in JP is approximately 15% by age 35 years and 68% by age 60 years (Jass et al., 1988; Murday et al., 1989). Juvenile polyps are primarily found in the colon and rectum but also in the small bowel and stomach and are characterized histologically by mucous retention, edema and granulation tissue (Rosai, 2011). Solitary juvenile polyps are sometimes found in the absence of JP (Nugent et al., 1993). Abdominal symptoms are usually noted in childhood or adolescent years. In a series of 145 cases, mean age at symptomatic onset was 6 years (Veale et al., 1966). In a subset of patients, JP is associated with hereditary hemorrhagic telangiectasia (Osler-Weber-Rendu syndrome; O'Malley et al., 2012). Surveillance is recommended for cancers of the colon, stomach and small bowel (Syngal et al., 2015).

Peutz-Jeghers syndrome

Peutz-Jeghers syndrome (PJS) is an autosomal dominant syndrome associated with CRC, Peutz-Jeghers polyps and mucocutaneous pigmentations. Its existence was first suggested by Jan Peutz in 1921 and confirmed by Harold Jeghers in 1949 (Jeghers et al., 1949). The reported incidence of PJS is 1:8,300-1:200,000 live births (Lier et al., 2010). During 1997-1998, the associated gene was localized to chromosome 19p, and mutations were detected in the serine-threonine kinase *STK11*, which was probably the first example of protein kinase inactivation underlying cancer susceptibility (Hemminki et al., 1997; Amos et al., 1997; Hemminki et al., 1998a). Histologically, Peutz-Jeghers polyps are characterized by smooth muscle fibers that protrude from the muscularis mucosae (Rosai, 2011). PJS may present in childhood with abdominal pain, intussusception, GI bleeding and rectal prolapse. In later years of life, the primary concern is predisposition to malignant tumors. Cancer risk in PJS is highly pleiotropic and includes cancers of the colorectum, pancreas, stomach, esophagus, small bowel, ovary, uterus, breast and lung (Giardiello et al., 2000). Accordingly, surveillance for multiple cancer types is recommended (Syngal et al., 2015).

PTEN hamartoma tumor syndrome

PTEN hamartoma tumor syndrome (PHTS) refers to a group of rare clinical syndromes (Cowden syndrome, Bannayan-Riley-Ruvalcaba syndrome, Proteus syndrome and Proteus-like syndrome) that are caused by germline mutations in *PTEN*, which is a negative regulator of the phosphoinositide-3-kinase pathway (Hobert et al., 2009). The mode of inheritance is autosomal dominant. Clinical manifestations are diverse and include intestinal and extraintestinal hamartomas and a spectrum of benign and malignant tumors. In a cohort of 127 *PTEN* mutation carriers, the standardized incidence ratio for CRC was 224 (95% CI, 119-403; Heald et al., 2010). PHTS is also associated with increased risks of breast, thyroid, endometrial and renal cancers and melanoma (Tan et al., 2012). Surveillance for multiple cancer types is recommended (Hobert et al., 2009).

Polymerase proofreading-associated polyposis

Germline mutations in the proofreading domains of *POLE* and *POLD1* were identified with the use of linkage analysis and WGS in families with suspected autosomal dominant CRC susceptibility, colorectal adenomas and endometrial cancer (Palles et al., 2013). Further studies have revealed recurrent *POLE* p.Leu424Val mutations in CRC patients with high-risk features (early age of onset, colorectal polyps, familial CRC or multiple primary CRCs; Spier et al., 2015; Elsayed et al., 2015; Bellido et al., 2016). The clinical characterization of *POLE*- and *POLD1*-related phenotypes is complicated by ascertainment bias in the published studies. Homologues of *POLE* p.Leu424Val and *POLD1* p.Ser478Asn cause polymerase infidelity and elevated mutation rates in yeast cells (Palles et al., 2013).

***NTHL1*-associated polyposis**

NTHL1-associated polyposis was detected in a WES study of 48 Dutch families with unexplained adenomatous polyposis. Seven individuals from three families were found to have biallelic LoF variants in *NTHL1* (p.Gln90Ter; MAF 0.144%; `http://gnomad.broadinstitute.org`), compatible with autosomal recessive inheritance (Weren et al., 2015). Biallelic loss of *NTHL1* has been reported in high-penetrance predisposition to multiple benign and malignant tumor types, but evidence remains scarce (Weren et al., 2015; Rivera et al., 2015). Similar to *MUTYH*, *NTHL1* is a component of the base-excision repair pathway, but mutations in these two genes are associated with different somatic mutation patterns (Weren et al., 2015).

Hereditary mixed polyposis

GREM1 has been implicated in autosomal dominant hereditary mixed polyposis syndrome in Ashkenazi Jewish families (Jaeger et al., 2012). The disease-associated locus was mapped to chromosome 15q13.3, and a 40-kb duplication was detected upstream of *GREM1.* The duplication was associated with increased allele-specific expression of *GREM1* (Jaeger et al., 2012).

3.3.3 Other hereditary syndromes

Li-Fraumeni syndrome

The risk of CRC in Li-Fraumeni syndrome is poorly understood. An association between early-onset CRC and classic Li-Fraumeni syndrome has been suggested (Wong et al., 2006). In a series of 457 early-onset CRC cases diagnosed before age 40 years, six patients (1.3%) had missense variants in *TP53* (Yurgelun et al., 2015). These six patients did not meet clinical criteria for Li-Fraumeni syndrome, and the pathogenicity of the identified variants could not be confirmed.

Oligodontia-CRC syndrome

AXIN2 has been implicated as a cause of autosomal dominant CRC susceptibility in rare families with variable numbers of colorectal adenomas with or without tooth agenesis (Lammi et al., 2004; Rivera et al., 2014). Based on these studies, *AXIN2* mutations are thought to be highly penetrant.

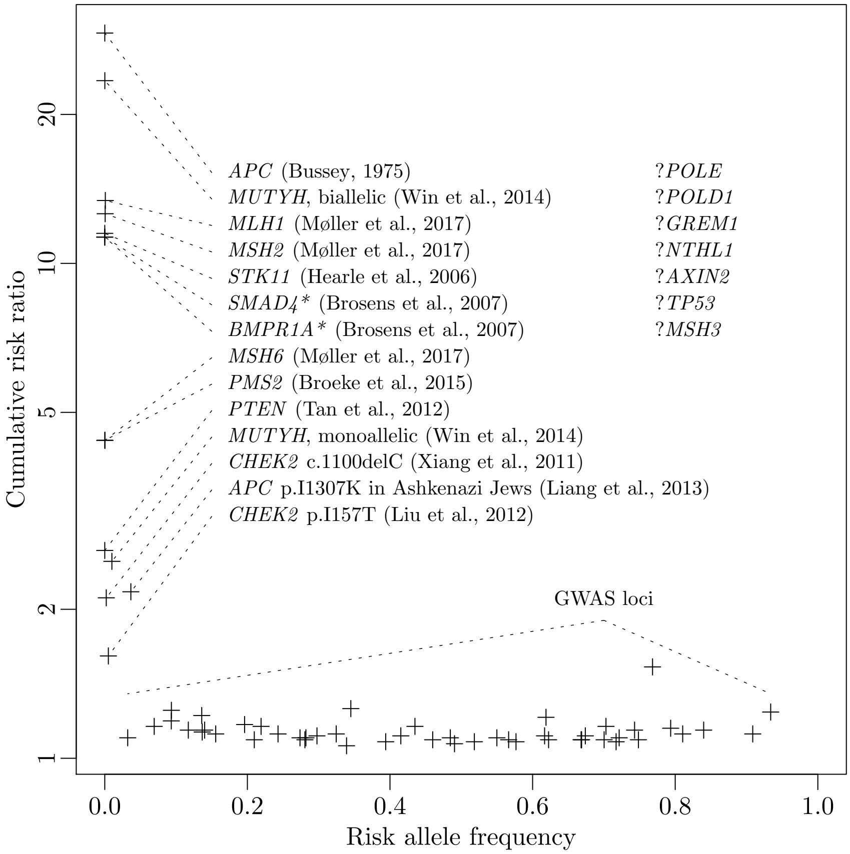


Figure 3.1: Genetic architecture of CRC susceptibility. Cumulative risk ratios were calculated from cumulative risks by age 70-80 years or ORs from case-control studies. The cumulative risk of CRC in the general population was assumed to be 3.4% by age 75 years (Ferlay et al., 2015). *Risk estimate for JP.

Table 3.1: Hereditary CRC syndromes. *MOI, mode of inheritance; AD, autosomal dominant; AR, autosomal recessive.*

Syndrome	**Chromosomal location(s)**	**Mutated gene(s)**	**MOI**	**Clinical features**
Lynch syndrome	3p22.2, 2p21, 2p16.3, 7p22.1, 2p21	*MLH1*, *MSH2*, *MSH6*, *PMS2*, *EPCAM*	AD	Few colorectal adenomas (usually <10); cancers of the endometrium, stomach, small bowel, pancreas, bile duct, kidney, urothelium, prostate, ovary and central nervous system
Constitutional mismatch repair deficiency syndrome	As in Lynch syndrome	As in Lynch syndrome	AR	Café-au-lait spots; childhood cancer (especially small bowel, brain, blood)
Familial adenomatous polyposis	5q22.2	*APC*	AD	GI adenomas ($\geq$100 in the colorectum); cancers of the brain, liver, thyroid, small bowel, pancreas and stomach; desmoids, osteoma, adrenal adenoma, retinal pigment epithelium hypertrophy
Attenuated familial adenomatous polyposis	As in familial adenomatous polyposis	5' or 3' end of *APC*	AD	GI adenomas ($<$100 in the colorectum); colonic and extracolonic manifestations more limited than in familial adenomatous polyposis
MUTYH-associated polyposis	1p34.1	*MUTYH*	AR	GI adenomas (typically $<$100 in the colorectum); duodenal cancer
Juvenile polyposis	18q21.2, 10q23.2	*SMAD4*, *BMPR1A*	AD	Juvenile polyps; pancreatic cancer, gastric cancer, duodenal cancer; hereditary hemorrhagic telangiectasia
Peutz-Jeghers syndrome	19p13.3	*STK11*	AD	Peutz-Jeghers polyps; mucocutaneous pigmentation; GI and gynecological cancers (small bowel, gastric, pancreatic, esophageal, ovarian, breast, endometrial)
PTEN hamartoma tumor syndrome	10q23.31	*PTEN*	AD	Multiple benign and malignant tumors; especially cancers of the breast, thyroid and endometrium; melanoma
Polymerase proofreading -associated polyposis	12q24.33, 19q13.33	*POLE*, *POLD1*	AD	Multiple and possibly large ($>$2 cm) colorectal adenomas; endometrial cancer
NTHL1-associated polyposis	16p13.3	*NTHL1*	AR	Multiple colorectal adenomas
Hereditary mixed polyposis	15q13.3	*GREM1*	AD	Colorectal polyps (atypical juvenile polyps, hyperplastic polyps and adenomas; total number usually $<$15)
Oligodontia-CRC syndrome	17q24.1	*AXIN2*	AD	Tooth agenesis; polyposis may or may not be present
Li-Fraumeni syndrome	17p13.1	*TP53*	AD	Breast cancer, sarcomas, brain tumors and leukemia

3.4 Moderate-penetrance variants

Moderate-penetrance variants are associated with two- to fourfold cancer risk (Sud et al., 2017). They are thought to be rare because of purifying selection (MAF $< 1\%$) but may become enriched in isolated populations in which genetic bottlenecks or founder effects have occurred (Hatzikotoulas et al., 2014; Gibson, 2012). To date, relatively few such variants have been convincingly associated with CRC.

APC p.Ile1307Lys is present in 7% of Ashkenazi Jewish individuals. In a meta-analysis, its OR for CRC was 2.17 (95% CI, 1.64-2.86) in Ashkenazi Jews, but the effect size was uncertain in other populations (Liang et al., 2013). This level of risk is comparable to having one first-degree relative with CRC, and screening recommendations are roughly similar (`https://www.nccn.org/`). At the nucleotide level, *APC* p.Ile1307Lys causes a hypermutable $(A)_8$ tract in exon 15 of *APC* (Laken et al., 1997).

Monoallelic *MUTYH* mutations may also be clinically significant. In the largest available meta-analysis, the RR for CRC by age 70 years was 2.5 (95% CI, 1.6-3.9), which suggests that monoallelic mutation carriers would benefit from more intensive screening than the general population (Win et al., 2014). Another meta-analysis, however, did not suggest a substantially increased risk (OR 1.14; 95% CI, 0.96-1.36), which makes the evidence somewhat inconsistent (Lubbe et al., 2009).

Meta-analyses suggest that *CHEK2* 1100delC and p.Ile157Tyr are associated with CRC risk. The ORs for 1100delC and p.Ile157Tyr have been estimated at 2.11 (95% CI, 1.41-3.16) and 1.61 (95% CI, 1.40-1.87), respectively (Liu et al., 2012; Xiang et al., 2011). However, the possibility of publication bias needs to be taken into account when interpreting these meta-analyses. Because *CHEK2* variants are associated with susceptibility to breast cancer and other cancer types (Näslund-Koch et al., 2016), it is a plausible candidate gene for CRC susceptibility.

3.5 Low-penetrance variants

To date, GWASs have identified at least 48 independent CRC risk SNPs (Table 3.2). These studies have been conducted in European, East Asian, African and Middle-Eastern populations. The establishment of international consortia has enabled meta-analyses of multiple primary GWASs (Tomlinson et al., 2010; Jia et al., 2013). Main results from published CRC GWASs are reviewed below and summarized in Table 3.2.

Tomlinson et al., 2007, genotyped 547,647 SNPs in 930 British individuals with familial CRC or high-risk adenoma and 960 matched cancer-free controls (the UK1 GWAS). Combined analysis of discovery and replication sets with 7,954 CRC cases and 6,206 controls showed a genome-wide significant association between rs6983267 (8q24.21) and CRC ($p = 1 \cdot 10^{-14}$). In a simultaneously published study, **Zanke et al., 2007,** genotyped 99 632 SNPs in 1,226 CRC patients and 1,239 matched controls from the Ontario Familial Colorectal Cancer Registry. After three replication stages, a genome-wide significant association was found between 8q24.21 (tagged by rs6983267 and rs10505477) and CRC. The CRC risk allele rs6983267-G has an increased binding affinity for the TCF4 transcription factor, which may lead to enhanced Wnt signaling (Tuupanen et al., 2009). In addition, the *MYC* oncogene is located within 350 kb from rs6983267 and may be a target of long-range regulatory interaction (Pomerantz et al., 2009). **Broderick et al., 2007,** replicated the second-ranked SNP in the UK1 GWAS, rs4939827 at 18q21.1 intronic to *SMAD7*; *SMAD7* was considered a plausible candidate gene for this association. In the UK2 GWAS, **Tomlinson et al., 2008,** genotyped 42,708 SNPs from the UK1 GWAS in 7,160 cases and 6,614 controls. Novel genome-wide significant associations were found at 8q23.3 (rs16892766 near *EIF3H*) and 10p14 (rs10795668 in an intergenic region).

Jaeger et al., 2008. A hereditary mixed polyposis locus had been previously mapped to 15q13.3-15q14 in Ashkenazi Jewish families, and it was hypothesized that this locus may also harbor common variants that predispose to CRC in the general population (Tomlinson et al., 1999; Jaeger et al., 2003). rs4779584 (15q13.3 between *SCG5* and *GREM1*) showed some evidence of association in the UK1 GWAS ($p = 4 \cdot 10^{-4}$), and genotyping of 7,961 CRC cases and 6,803 controls confirmed a clear association between rs4779584 and CRC ($p = 4 \cdot 10^{-14}$). *GREM1* is an antagonist of TGF-β signaling that is often expressed in cancer-associated stromal cells (Sneddon et al., 2006).

Tenesa et al., 2008, genotyped 541,628 SNPs in 981 Scottish early-onset

CRC cases (< 55 years at diagnosis) and 1,002 matched controls (the Scotland1 GWAS). Meta-analysis of 17,457 cases and 16,353 controls of different ancestries revealed a novel genome-wide significant association at 11q23.1 (rs3802842, intronic to *COLCA1* and *COLCA2*). rs3802842 (11q23.1) and the previously reported SNP rs4939827 (18q21.1) showed site-specificity for rectal cancer ($p < 0.01$).

COGENT Study, 2008. The COGENT consortium undertook meta-analysis of 38,710 SNPs in four studies: UK1, UK2, Scotland1 and Scotland2. The total sample size, including replication sets, was 20,186 cases and 20,855 controls. Four novel loci showed genome-wide significant associations with CRC: 14q22.2 (rs4444235 near *BMP4*), 16q22.1 (rs9929218 intronic to *CDH1*), 19q13.1 (rs10411210 intronic to *RHPN2*) and 20q12.3 (rs961253 near *CASC20*). rs9929218 is located in intron 1 of *CDH1*. Because mutations in *CDH1* are associated with hereditary diffuse gastric cancer (Guilford et al., 1998), it was hypothesized that *CDH1*-related cancer susceptibility may involve a spectrum of both common and rare alleles. rs4444235 (upstream of *BMP4*) showed specificity for MSS CRC ($p = 2 \cdot 10^{-3}$).

Houlston et al., 2010. CRC patients from the VICTOR and QUASAR2 RCTs (n = 1,432) and controls from the UK 1958 Birth Cohort (n = 2,697) were combined to form the VQ58 GWAS. VQ58 was meta-analyzed with the UK1, UK2, Scotland1 and Scotland2 GWASs. Four new CRC susceptibility loci were identified: 1q41 (rs6691170 near *LINC02257*), 3q26.2 (rs10936599, a synonymous coding variant in *MYNN*), 12q13.12 (rs11169552 near *ATF1*) and 20q13.33 (rs4925386, intronic to *LAMA5*).

Because the bone morphogenetic protein signaling pathway had been previously implicated in CRC susceptibility, **Tomlinson et al., 2011,** hypothesized that multiple independent CRC-associated SNPs may exist at 14q22.2 (near *BMP4*), 15q13.3 (near *SCG5* and *GREM1*) and 20p12.3 (near *BMP2*). In an analysis of 24,910 CRC cases and 26,275 controls of European ancestry, rs1957636 at 14q22.2 (near *BMP4*) and rs4813802 at 20p12.3 (near *BMP2*) showed strong associations with CRC despite weak LD with the originally reported tag SNPs rs4444235 (14q22.2) and rs961253 (20p12.3), respectively.

Dunlop et al., 2012, meta-analyzed five previously reported GWASs (Scotland1, Scotland2, UK1, UK2 and VQ58) and 260 promising SNPs in the COIN and COIN-B RCTs (Maughan et al., 2011; Wasan et al., 2014). Three new loci were genome-wide significant in combined analysis of 29,778 cases and 29,204 controls of European ancestry: 6p21.2 (rs1321311 near *PANDAR* and *CDKN1A*), 11q13.4 (rs3824999, intronic to *POLD3*) and Xp22.2 (rs5934683 near *GPR143* and *SHROOM2*). A plausible target gene for rs1321311 (6p21.2) is *CDKN1A*, which encodes p21, a protein that mediates p53-dependent cell cycle arrest and may have p53-independent tumor suppressor functions (Abbas et al., 2009). The risk allele of rs5934683 (Xp22.2) was associated with downregulation of *SHROOM2* in both normal and neoplastic colorectal tissue ($p = 1 \cdot 10^{-7}$).

Jia et al., 2013, reported a GWAS in collaboration with the Asian Colorectal Cancer Consortium. In the discovery stage, 1,636,380 SNPs were imputed and analyzed in 2,098 CRC cases and 5,749 controls from China, South Korea and Japan. In joint analysis of East Asian and European datasets, three new loci were significantly associated with CRC: 5q31.1 (rs64716, intronic to *C5orf66* and near *PITX1*), 20p12.3 (rs2423279 near *HAO1*) and 12p13.32 (rs10774214 near *CCND2* and *CCND2-AS1*). *CCND2* encodes cyclin D2, which is related to the cyclin D1 proto-oncogene (Kim et al., 2009).

Zhang et al., 2014b, imputed and analyzed over 2 million SNPs in 2,038 CRC cases and 6,172 cancer-free controls from China, Japan and South Korea. The four-stage design comprised a total of 14,963 CRC cases and 31,945 controls. Sixteen genome-wide significant loci were identified, and among these were six novel CRC risk loci: 10q22.3 (rs704017 within *ZMIZ-AS1*), 10q25.2 (rs11196172, intronic to *TCF7L2*), 11q12.2 (rs174537 within *MYRF* and *TMEM258*), 12p13.31 (rs10849432 near *CD9*), 17p13.3 (rs12603526, intronic to *NXN*) and 19q13.2 (rs1800469, intronic to *B9D2*). The risk allele rs11196172-A was associated with increased expression of both *TCF7L2* and *VTI1A* in colonic tumor tissue ($p = 0.003$). rs1800469 maps to the promoter region of *TGFB1*, and the risk allele (G) has been associated with lower expression and circulating levels of *TGFB1* (Grainger et al., 1999).

Wang et al., 2014, conducted a meta-analysis of 2,627 cases and 3,797 controls of Japanese ancestry and 1,894 cases and 4,703 controls of African American ancestry. In the discovery meta-analysis, rs12241008 in intron 3 of *VTI1A* near *TCF7L2* (10q25.2) showed a novel genome-wide significant association with CRC ($p = 2.9 \cdot 10^{-8}$). The association was also genome-wide significant in combined analysis with European studies with 21,344 cases and 26,711 controls ($p = 1.5 \cdot 10^{-9}$).

Zhang et al., 2014a, imputed and tested 1,695,815 SNPs in 1,773 CRC cases and 2,642 controls of East Asian ancestry. Combined analysis with replication data (8,675 cases and 10,504 controls) revealed a new genome-wide significant association between rs7229639 (18q21.1, intron 3 of *SMAD7*) and CRC. The association was independent of the previously reported risk SNP rs4939827, which is located 2.5 kb downstream from rs7229639.

Whiffin et al., 2014, meta-analyzed five European-ancestry GWASs (UK1, Scotland1, VQ58, CCFR1 and CCFR2) that comprised 5,626 cases and 7,817 controls. Replication was undertaken in additional 14,037 cases and 15,937 controls. 10q24.2 (rs1035209 near *SLC25A28*) was implicated as a novel CRC risk locus. In addition, two loci that had been previously suggested by Peters et al., 2013 reached genome-wide significance: 1q25.3 (rs10911251, intronic to *LAMC1*) and 12p13.32 (rs3217810, intronic to *CCND2*) (Peters et al., 2013).

Schmit et al., 2014, meta-analyzed CRC GWASs from Israel (1,616 cases and 1,329 controls) and the CCFR (1,977 cases and 999 controls). Results were replicated in 1,131 cases and 831 controls. In combined analysis, rs17042479 (4q32.2 near *FSTL5*) was associated with CRC with a relatively large effect size (OR, 1.53; 95% CI, 1.39-1.67).

Al-Tassan et al., 2015, studied 2,244 cases from the COIN and COIN-B RCTs and 2,674 controls from the UK Blood Service Control Group (COIN GWAS). The COIN GWAS was meta-analyzed with the UK1, Scotland1, VQ58, CCFR1 and CCFR2 GWASs. Genotypes for over 10 million variants were imputed in 7,577 cases and 9,979 controls. Eight previously reported loci had a p-value $< 5 \cdot 10^{-8}$, and a novel genome-wide significant association was found at 1p36.12 (rs72647484 near *WNT4*, *CDC42* and *MIR4418*). There was a borderline significant ($p = 5.06 \cdot 10^{-8}$) association at 16q24.1 (rs16941835 within *AC009154.1*). *WNT4* has been shown to activate the canonical Wnt-β-catenin signaling pathway (Lyons et al., 2004).

Schumacher et al., 2015, meta-analyzed 19 studies that comprised 18,299 CRC cases and 19,656 controls from four consortia. Results were replicated in Asian populations with 4,725 cases and 9,969 controls, and six new CRC risk loci were identified: 3p22.1 (rs35360328 near *CTNNB1 AC099560.1*, 3p14.1 (rs812481, intronic to *LRIG1*), 10q24.2 (rs11190164 near *SLC25AA28*), 12q24.12 (rs3184504, *SH2B3* p.Trp262Arg), 12q24.22 (rs73208120, intronic to *NOS1*) and rs6066825 (20q13.13, intronic to *PREX1*). rs35360328 is located 320 kb upstream from *CTNNB1*, pointing to β-catenin regulation as a candidate mechanism for the association.

Cheng et al., 2015, hypothesized that common variants may be associated with both colorectal and endometrial cancers, similar to germline mutations in DNA repair genes. Sixteen datasets with 13,265 cases and 40,245 controls were analyzed. 12q24.12 (*SH2B3* p.Trp262Arg) was associated with the composite phenotype of colorectal or endometrial cancer ($p = 7.23 \cdot 10^{-9}$). In addition, 18q22 (rs12970291 near *TSHZ1*) was associated with CRC and endometrial cancer with opposing effects on the two diseases ($p = 4.82 \cdot 10^{-8}$).

Wang et al., 2016, genotyped 691,326 SNPs in 1,023 CRC cases and 1,306 controls of Han Chinese ancestry. Combined analysis of discovery and replication sets with 5,317 cases and 6,887 controls implicated rs2238126 (12p13.2, intron 4 of *ETV6*) as a novel CRC susceptibility variant.

Zeng et al., 2016, conducted one of the largest GWASs of CRC in East Asians to date. The discovery set comprised 8,027 CRC cases and 22,577 controls from China, Japan and South Korea. A total of 11,044 CRC cases and 12,047 controls of Asian ancestry were studied in the replication stage, which led to the identification of four novel CRC risk loci: 6p21.1 (rs4711689, intronic to *TFEB*), 8q23.3 (rs2450115 near *EIF3H*), 10q24.3 (rs4919687 within *CYP17A1*) and 12p13.3 (rs11064437 in a splice acceptor site of *SPSB2*). rs2450115 is located within 7 kb from the previously reported CRC risk SNP rs16892766 (Tomlinson et al., 2008), but these variants are essentially independent ($r^2 < 0.05$ in both Asian and European populations).

Orlando et al., 2016, meta-analyzed six previous GWASs (UK1, Scotland1, VQ58, CCFR1, CCFR2 and COIN) with a Finnish GWAS with 1,172 cases and 8,266 controls. Over 10 million variants were imputed and analyzed in 13,656 cases and 21,667 controls. Ten previously reported SNPs were genome-wide significant, and a suggestive association ($p < 1 \cdot 10^{-5}$) was found at a novel locus at 2q35 (rs992157, intronic to *PNKD* and *TMBIM1*). After replication in 5,061 cases and 3,509 controls of European ancestry ($p = 0.023$), combined analysis of 18,717 cases and 25,176 controls showed a genome-wide significant association between rs992157 and CRC ($p = 3.15 \cdot 10^{-8}$). Since rs992157 is in strong LD with the IBD risk SNP rs2382817 ($r^2 = 0.90$), other IBD risk SNPs were also tested, and 11 additional IBD risk SNPs were associated with CRC at $q < 0.05$. rs992157 and rs2382817 show opposing effects on CRC and IBD risk, which is consistent with a pleiotropic rather than IBD-mediated effect on CRC risk.

Table 3.2: Common risk alleles for CRC and nearest genes by date of publication. Allele frequencies were obtained from the gnomAD browser (`http://gnomad.broadinstitute.org`, accessed in November 2018). *RAF, risk allele frequency.*

Chr.	**SNP**	**RAF**	**Gene**	p**-value**	**OR [95% CI]**	**Reference**
8q21.24	rs6983267-G	0.62	*CASC8*	$1 \cdot 10^{-14}$	1.21 [1.15, 1.27]	Tomlinson et al., 2007
18q21.1	rs4939827-T	0.44	*SMAD7*	$1 \cdot 10^{-12}$	1.16 [1.09, 1.27]	Broderick et al., 2007
8q23.3	rs16892766-C	0.093	*EIF3H*	$3 \cdot 10^{-18}$	1.25 [1.19, 1.32]	Tomlinson et al., 2008
10p14	rs10795668-A	0.24	NA	$3 \cdot 10^{-13}$	1.12 [1.10, 1.16]	Tomlinson et al., 2008
15q13.3	rs4779584-T	0.35	*SCG5*	$4 \cdot 10^{-14}$	1.26 [1.19, 1.34]	Jaeger et al., 2008
11q23.1	rs3802842-C	0.30	*COLCA1*	$6 \cdot 10^{-10}$	1.11 [1.08, 1.15]	Tenesa et al., 2008
14q22.2	rs4444235-C	0.42	*BMP4*	$8 \cdot 10^{-10}$	1.11 [1.08, 1.15]	COGENT Study, 2008
16q22.1	rs9929218-G	0.72	*CDH1*	$1 \cdot 10^{-8}$	1.10 [1.06, 1.12]	COGENT Study, 2008
19q13.1	rs10411210-C	0.79	*RHPN2*	$5 \cdot 10^{-9}$	1.15 [1.10, 1.20]	COGENT Study, 2008
20p12.3	rs961253-A	0.32	*CASC20*	$2 \cdot 10^{-10}$	1.12 [1.08, 1.16]	COGENT Study, 2008
1q41	rs6691170-T	0.34	*LINC02257*	$1 \cdot 10^{-9}$	1.06 [1.03, 1.09]	Houlston et al., 2010
3q26.2	rs10936599-C	0.72	*MYNN*	$3 \cdot 10^{-8}$	1.08 [1.04, 1.10]	Houlston et al., 2010
12q13.12	rs11169552-C	0.75	*ATF1*	$2 \cdot 10^{-10}$	1.09 [1.05, 1.11]	Houlston et al., 2010
20q13.33	rs4925386-C	0.58	*LAMA5*	$2 \cdot 10^{-10}$	1.08 [1.05, 1.10]	Houlston et al., 2010
14q22.2	rs1957636-A	0.48	*BMP4*	$4 \cdot 10^{-10}$	1.08 [1.06, 1.11]	Tomlinson et al., 2011
20p12.3	rs4813802-G	0.28	*LINC01713*	$5 \cdot 10^{-11}$	1.09 [1.06, 1.12]	Tomlinson et al., 2011
6p21.2	rs1321311-A	0.27	*PANDAR*	$1 \cdot 10^{-10}$	1.10 [1.07, 1.13]	Dunlop et al., 2012
11q13.4	rs3824999-G	0.39	*POLD3*	$4 \cdot 10^{-10}$	1.08 [1.05, 1.10]	Dunlop et al., 2012
Xp22.2	rs5934683-T	0.49	*GPR143*	$7 \cdot 10^{-10}$	1.07 [1.04, 1.10]	Dunlop et al., 2012
5q31.1	rs647161-A	0.62	*C5orf66*	$1 \cdot 10^{-10}$	1.11 [1.08, 1.15]	Jia et al., 2013
12p13.32	rs2423279-C	0.28	*HAO1*	$7 \cdot 10^{-9}$	1.10 [1.06, 1.14]	Jia et al., 2013
20p12.3	rs10774214-T	0.46	*CCND2-AS1*	$3 \cdot 10^{-8}$	1.09 [1.06, 1.13]	Jia et al., 2013
10q22.3	rs704017-G	0.55	*ZMIZ1-AS1*	$2 \cdot 10^{-8}$	1.10 [1.06, 1.13]	Zhang et al., 2014b
10q25.2	rs11196172-A	0.14	*TCF7L2*	$1 \cdot 10^{-12}$	1.14 [1.10, 1.18]	Zhang et al., 2014b
11q12.2	rs174537-G	0.70	*MYRF*	$9 \cdot 10^{-21}$	1.16 [1.12, 1.19]	Zhang et al., 2014b
12p13.31	rs10849432-T	0.84	*CD9*	$6 \cdot 10^{-10}$	1.14 [1.09, 1.18]	Zhang et al., 2014b
17p13.3	rs12603526-C	0.032	*NXN*	$3 \cdot 10^{-8}$	1.10 [1.06, 1.14]	Zhang et al., 2014b
19q13.2	rs1800469-G	0.70	*B9D2*	$1 \cdot 10^{-8}$	1.09 [1.06, 1.12]	Zhang et al., 2014b
10q25.2	rs12241008-C	0.14	*VTI1A*	$1 \cdot 10^{-9}$	1.13 [1.09, 1.18]	Wang et al., 2014
18q21.1	rs7229639-A	0.14	*SMAD7*	$3 \cdot 10^{-11}$	1.22 [1.15, 1.29]	Zhang et al., 2014a
1q25.3	rs10911251-A	0.62	*LAMC1*	$2 \cdot 10^{-8}$	1.09 [1.06, 1.12]	Whiffin et al., 2014
10q24.2	rs1035209-T	0.16	*SLC25A28*	$5 \cdot 10^{-11}$	1.12 [1.08, 1.16]	Whiffin et al., 2014
12p13.32	rs3217810-T	0.093	*CCND2*	$2 \cdot 10^{-10}$	1.19 [1.13, 1.25]	Whiffin et al., 2014
4q32.2	rs17042479-A	0.77	*FSTL5*	$8 \cdot 10^{-9}$	1.53 [1.39, 1.67]	Schmit et al., 2014
1p36.12	rs72647484-T	0.93	*MIR4418*	$1 \cdot 10^{-8}$	1.24 [1.15, 1.33]	Al-Tassan et al., 2015
16q24.1	rs16941835-C	0.22	*AC009154.1*	$5 \cdot 10^{-8}$	1.16 [1.09, 1.22]	Al-Tassan et al., 2015
3p22.1	rs35360328-A	0.12	*AC099560.1*	$3 \cdot 10^{-9}$	1.14 [1.09, 1.19]	Schumacher et al., 2015
3p14.1	rs812481-G	0.67	*LRIG1*	$2 \cdot 10^{-8}$	1.09 [1.05, 1.12]	Schumacher et al., 2015
10q24.2	rs11190164-G	0.21	*NKX2-3*	$4 \cdot 10^{-8}$	1.09 [1.06, 1.12]	Schumacher et al., 2015
12q24.12	rs3184504-C	0.67	*SH2B3*	$2 \cdot 10^{-8}$	1.09 [1.06, 1.12]	Schumacher et al., 2015
12q24.22	rs73208120-G	0.069	*NOS1*	$3 \cdot 10^{-8}$	1.16 [1.11, 1.23]	Schumacher et al., 2015
20q13.13	rs6066825-A	0.57	*PREX1*	$4 \cdot 10^{-9}$	1.09 [1.06, 1.12]	Schumacher et al., 2015
12p13.2	rs2238126-G	0.20	*ETV6*	$3 \cdot 10^{-10}$	1.17 [1.11, 1.23]	Wang et al., 2016
6p21.1	rs4711689-A	0.67	*TFEB*	$4 \cdot 10^{-8}$	1.11 [1.07, 1.15]	Zeng et al., 2016
8q23.3	rs2450115-T	0.81	*EIF3H*	$1 \cdot 10^{-12}$	1.12 [1.09, 1.15]	Zeng et al., 2016
10q24.3	rs4919687-G	0.74	*CYP17A1*	$8 \cdot 10^{-12}$	1.14 [1.10, 1.19]	Zeng et al., 2016
12p13.3	rs11064437-C	0.91	*SPSB2*	$4 \cdot 10^{-11}$	1.12 [1.08, 1.16]	Zeng et al., 2016
2q35	rs992157-A	0.49	*PNKD*	$3 \cdot 10^{-8}$	1.10 [1.06, 1.13]	Orlando et al., 2016

Aims of the study

1. Hereditary cancer syndromes are frequent in patients with early-onset CRC, but the optimal diagnostic strategy is unknown. The first aim of the study was to assess the practical feasibility and added value of WES in diagnosing hereditary cancer syndromes in early-onset CRC patients.

2. Rare genetic variants with moderate-to-high penetrance are clinically and biologically significant and may explain part of the missing heritability of CRC. Therefore, the second aim was to study rare protein-coding variants in Finnish early-onset and familial CRC patients with unexplained etiology.

3. GWASs of CRC have been successful in identifying common SNPs with high population attributable risk and substantial cumulative effects. Also, large haplotype reference panels have recently become available, improving the performance of genotype imputation. The third aim was to conduct a GWAS to identify common and low-frequency SNPs that influence CRC risk in the Finnish population and other populations of European ancestry.

Materials and methods

4 Study participants (I, II, III)

4.1 The Finnish CRC Collection (I, II, III)

Studies I, II and III were conducted in accordance with the Declaration of Helsinki and approved by the Finnish National Supervisory Authority for Welfare and Health, National Institute for Health and Welfare (THL/151/5.05.00 /2017) and the Ethics Committee of the Hospital District of Helsinki and Uusimaa (HUS/408/13/03/03/09). All participants provided written informed consent. Nine Finnish central hospitals recruited 1,042 CRC patients for the Finnish CRC Collection between May 1994 and June 1998 (the C series). Tumor samples from these 1,042 patients had been screened for MSI as described previously (Aaltonen et al., 1998; Salovaara et al., 2000), and *MLH1* and *MSH2* had been screened for germline mutations in patients with MSI CRC. After June 1998, sample collection was continued in two of the nine central hospitals (the S series). Patients recruited after June 1998 were tested for tumor MSI with the Bethesda microsatellite panel (BAT25, BAT26, D5S346, D17S250 and D2S123; Boland et al., 1998), and data on germline mutations in MMR genes were obtained from diagnostic laboratories. In studies I and II, we analyzed 1,042 patients from the C series and the 472 patients from the S series (total, 1,514; 50% male; median age at diagnosis, 68 years; IQR, 59-76 years). Of the 1,514 CRC cases, 38 (2.5%) had been diagnosed before age 40 years. In stage 1 of study III, termed the FIN GWAS, we analyzed 765 patients from the C series and 845 patients from the S series (total, 1,610; 53% male; median age at diagnosis, 69 years; IQR, 60-76 years). Medical records were reviewed for relevant clinical phenotypes and results of genetic testing. Data on first-degree relatives and their cancer diagnoses were acquired from national population registries and the Finnish Cancer Registry. The Finnish Cancer Registry has nearly complete coverage of incident cancers in Finland beginning from 1953 (Pukkala et al., 2018).

4.2 The FINRISK Study (III)

The FINRISK Study was initiated in 1972 to study the risk factors of major noncommunicable diseases in the Finnish population (Borodulin et al., 2017). Independent population surveys, including biological sample collection, have been carried out every five years beginning in 1972. DNA samples were first collected in year 1992. The FINRISK Study contributed 91 CRC cases and 14,187 cancer-free controls to the FIN GWAS (stage 1 of study III). These individuals had participated in sample collection in 1992, 1997, 2002 or 2007. In stage 2 of study III, we performed targeted genotyping in a smaller series of FINRISK Study participants who had not been included in stage 1 due to unavailable SNP array data (198 CRC cases and 172 cancer-free controls; referred to as the FINRISK series in stage 2 of study III). Data on cancer diagnoses in the FINRISK Study participants had been obtained from the Finnish Cancer Registry, and DNA samples were provided by the THL Biobank, Finland (`https://www.thl.fi/fi/web/thlfi-en/topics/information-packages/thl-biobank`).

4.3 Nordic studies (III)

Nordic studies from Sweden (STHLM2 and Gothenburg), Norway (HUNT) and Estonia contributed samples to study III. The STHLM2 series consisted of men who had been referred to prostate-specific antigen screening in Stockholm County, Sweden, between 2010 and 2012. DNA samples from the STHLM2 series were provided by the Karolinska Institute Biobank (`http://ki.se/forskning/ki-biobank`). The Gothenburg series consisted of cases and controls from the Sahlgrenska University Hospital, Gothenburg, Sweden. DNA samples from the Gothenburg series were provided by the Sahlgrenska Biobank (`https://www.gothiaforum.com/sab`). The HUNT series consisted of sample donors from the Nord-Trøndelag Health Study (HUNT) and Biobank (`https://www.ntnu.edu/hunt`). The Estonia series consisted of sample donors of the Estonian Genome Center Biobank (`www.geenivaramu.ee/en`). When possible, cases and controls were matched by year of birth and sex.

5 Methods (I, II, III)

5.1 Exome sequencing (I, II)

Twenty-three of the 38 early-onset CRC patients had unexplained etiology and underwent WES. Protein-coding DNA sequences were captured with SureSelect Human All Exon Kit v.1 (Agilent Technologies, Santa Clara, California). Paired-end 75 base-pair reads were generated with Illumina Genome Analyzer II (Illumina, San Diego, California). FASTQC was used for quality control of the raw WES data (https://www.bioinformatics.babraham.ac.uk/projects/fastqc/). The Burrows-Wheeler Aligner was used to map reads to the Genome Reference Consortium reference genome 37 (https://www.ncbi.nlm.nih.gov/assembly/GCF_000001405.13/). Duplicate reads were removed with Picard MarkDuplicates (http://broadinstitute.github.io/picard/). Genome Analysis Toolkit IndelRealigner (Broad Institute, Cambridge, MA) was used for local read realignment. Variants were called with Genome Analysis Toolkit UnifiedGenotyper (v.2.2-16-g9f648cb). Eighty-seven percent of the targeted regions were covered by more than 10 reads, and the mean coverage was 54. BasePlayer was used to visualize and annotate WES data (Katainen et al., 2017). In study I, we excluded variants that were detected in 212 WES controls (68 in-house controls or 144 Finnish migraine patients) or the 1,000 Genomes Project Phase 1 release (www.1000genomes.org) as many them were considered to represent common polymorphisms or sequencing artefacts. When relevant, reads were inspected to exclude false-positive variant calls.

To uncover new candidate genes for CRC susceptibility, we searched for rare protein-coding variants in 22 early-onset CRC patients without known predisposing conditions. An independent series of 95 genetically unexplained familial CRC patients, also of Finnish ancestry, were studied as a validation set (Gylfe et al., 2013). In order to prioritize variants that were most likely to be pathogenic, we focused on variants with MAF $< 0.1\%$ (Kryukov et al., 2007; MacArthur et al., 2012). In particular, it was hypothesized that LoF (nonsense, splice site or frameshift) variants may contribute to CRC susceptibility. Although many genetic disorders are caused by rare LoF variants, healthy humans carry approximately 100 LoF variants, which complicates the assessment of their pathogenicity (MacArthur et al., 2012). Loss-of-Function Transcript Effect Estimator (LOFTEE; `https://github.com/konradjk/loftee`) was used to exclude protein-truncating variants that were unlikely to abolish gene function (ancestral LoF alleles, LoF variants located in the last 5% of the coding region, splice site variants in small introns (<15 base pairs) and LoF variants surrounded by non-canonical splice sites). Functional effects of missense variants were predicted with PolyPhen-2 and SIFT (`http://www.ensembl.org`). Allele frequencies were determined in 3,374 Finnish and 58,112 non-Finnish control exomes that were publicly available in the Exome Aggregation Consortium (ExAC) database (Cambridge, MA; `http://exac.broadinstitute.org`, accessed in November 2014). PolyPhen-2 (`http://genetics.bwh.harvard.edu/pph2/`) and SIFT (`http://sift.jcvi.org`) were used to predict functional effects of missense variants.

5.2 Sanger sequencing (I, II)

When relevant, WES variant calls were validated by Sanger sequencing. AmpliTaqGold DNA polymerase (Applied Biosystems, Foster City, CA) was used in PCR reactions. PCR products were purified with ExoSAP-IT reagent (USB Corporation, Cleveland, OH, USA). Big Dye Terminator v3.1 chemistry (Applied Biosystems, Foster City, CA) was used for DNA sequencing. Applied Biosystems 3730xl DNA analyzer was used for gel electrophoresis. Allelic imbalance was assessed by comparing allele peak heights in normal and tumor DNA samples as described previously (Tuupanen et al., 2008).

5.3 SNP array genotyping (III)

To identify common and low-frequency SNPs that influence CRC risk in the general population, we conducted a GWAS in 1,701 Finnish CRC cases and 14,082 cancer-free controls (stage 1 of study III; the FIN GWAS). The Haplotype Reference Consortium panel was used to impute genotypes across a wide range of allele frequencies (McCarthy et al., 2016). Most promising SNPs were genotyped in 4,070 Nordic CRC cases and 2,377 controls (stage 2) and analyzed in previously published GWASs of CRC comprising 7,577 cases and 9,979 controls of European ancestry (stage 3) (Al-Tassan et al., 2015). Finally, data from the three stages were combined, and 13,348 cases and 26,438 controls were meta-analyzed. The study scheme is shown in Figure 5.1.

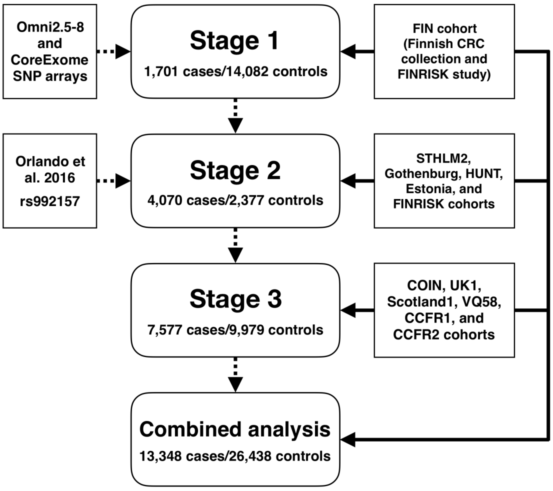

Omni2.5-8 and CoreExome SNP arrays
Stage 1
1,701 cases/14,082 controls
FIN cohort (Finnish CRC collection and FINRISK study)
Orlando et al. 2016
rs992157
Stage 2
4,070 cases/2,377 controls
STHLM2, Gothenburg, HUNT, Estonia, and FINRISK cohorts
Stage 3
7,577 cases/9,979 controls
COIN, UK1, Scotland1, VQ58, CCFR1, and CCFR2 cohorts
Combined analysis
13,348 cases/26,438 controls

Figure 5.1: Scheme of study III. The flow of genetic variants is shown as dotted arrows and the flow of samples as solid arrows.

In the Finnish CRC Collection, DNA samples from normal colorectal tissue or blood were genotyped with the Illumina (San Diego, CA) HumanOmni2.5-8 SNP array. In the FINRISK Study, blood DNA samples were genotyped with the Illumina HumanCoreExome SNP array. The MassARRAY System by Agena Bioscience (San Diego, CA) was used to genotype Nordic cases and controls (STHLM2, 544 cases/541 controls; Gothenburg, 1,903 cases/258 controls; HUNT, 1,168 cases/1,147 controls; Estonia, 257 cases/259 controls; FINRISK, 198 cases/172 controls), as well as 1,038 individuals from stage 1 for quality control purposes (925 individuals also genotyped with the HumanOmni2.5-8 array and 113 individuals also genotyped with the HumanCoreExome array).

PLINK v.1.90b3i was used for quality control of the SNP array data (`www.cog-genomics.org/plink/1.9/`). A total of 122 samples (17 genotyped with the HumanOmni2.5-8 array and 105 genotyped with the HumanCoreExome array) were excluded because of close relatedness (identity-by-descent coefficient > 0.2), sample duplication, discordant sex information or low genotyping rate. The remaining 1,701 CRC cases and 14,082 cancer-free controls were included in the FIN GWAS. The HumanOmni2.5-8 SNP array contained 2,315,673 autosomal sites, 273,074 of which were also found on the HumanCoreExome SNP array (`https://support.illumina.com/downloads.html`). Low-quality SNPs were excluded on the basis of low genotyping rate ($< 95\%$), excess homozygosity (rare homozygote frequency exceeding the heterozygote frequency, or any rare homozygous genotype with minor allele frequency (MAF) $< 2\%$), deviation from the Hardy-Weinberg equilibrium ($p < 1 \cdot 10^{-8}$), differential missingness between genotyping batches ($p < 1 \cdot 10^{-8}$), differential LD patterns between cases and controls or LD-based strand inconsistency.

After quality control, 214,705 SNPs were phased with SHAPEIT v2 (r790). The Haplotype Reference Consortium reference panel was used to impute genotypes (McCarthy et al., 2016). After genotype imputation, we excluded variants with low allele frequency ($< 0.4\%$) or low IMPUTE2 info score (< 0.4).

5.4 Association analysis (III)

The primary analysis in the FIN GWAS was based on a linear mixed model adjusted for age and sex (BOLT-LMM-inf; `https://data.broadinstitute.org/alkesgroup/BOLT-LMM/`). The age covariate was defined as age at CRC diagnosis in cases and age at end of follow-up in controls. Genetic effects were assumed to be additive. The genomic inflation factor was estimated by dividing the observed median of the BOLT-LMM-inf test statistic by the median of the χ^2_1 distribution. PLINK v.1.90b3i was used for LD-based SNP pruning and principal component analysis. Principal component analysis was performed with 13,012 LD-pruned SNPs with allele frequency $> 5\%$ and IMPUTE2 info score > 0.9. To assess imputation accuracy, we calculated squared Pearson correlation coefficients (r^2) between IMPUTE2 genotype dosage and MassARRAY genotype.

To generate uniform data for meta-analysis, the FIN series was reanalyzed by unconditional logistic regression under an additive genetic model, adjusting for sex, log-transformed age and 10 principal components with SNPTEST v.2.5.2. In the MassARRAY-genotyped Nordic datasets, unconditional logistic regression was applied with a minimum allele count of 10 using R v.3.3.3. Meta-analysis was performed with the metafor package v.1.9-9 in R v.3.3.3. Genomic control was applied by multiplying the standard errors of regression coefficients by the square root of the study-specific genomic inflation factor. Estimates of log ORs and standard errors were combined to obtain summary p-values, ORs and 95% CIs under inverse-variance weighted random-effects and fixed-effect models (function rma.uni in the metafor package). The Benjamini-Hochberg method was used to control the false discovery rate. The Bonferroni correction was applied when relevant. The type 1 error rate (α) was 0.05, and the genome-wide significance threshold was $5 \cdot 10^{-8}$. All p-values were two-sided.

Results

6 Exome sequencing of early-onset colorectal cancer patients

6.1 Hereditary syndromes in early-onset colorectal cancer patients (I)

Of 1,514 unselected Finnish CRC cases, 38 (2.5%) had been diagnosed before age 40 years. Among these 38 cases, the median age at diagnosis was 35 years (range, 21 to 39 years). Twenty of the 38 patients (53%) were male. All 38 cases were tested for tumor MSI, and 14 tumors (37%) were found to be microsatellite-unstable. Eleven of the 14 patients with MSI CRC (79%) had been screened for germline mutations in *MLH1* and *MSH2*. GI polyposis syndromes had been diagnosed clinically in four patients (three patients with FAP and one patient with JP). Testing for tumor MSI and evaluation for GI polyposis had led to the identification of hereditary CRC syndromes in 15 of the 38 patients (39%; Table 6.2). To assess the added value of WES as a diagnostic test, the remaining 23 patients were exome sequenced. WES data were analyzed for nonsynonymous variants in 10 genes that had been implicated in high-penetrance CRC predisposition (*MLH1*, *MSH2*, *MSH6*, *PMS2*, *APC*, *MUTYH*, *SMAD4*, *BMPR1A*, *STK11* and *PTEN*).

6.1.1 Added diagnostic value of exome sequencing

WES identified a splice-site mutation in *MLH1* (c.454-1G>A; Finnish founder mutation 2) in a 35-year-old patient with MSI CRC (s171). The mutation was confirmed by Sanger sequencing. *MLH1* c.454-1G>A could have been detected by MSI testing and subsequent Sanger sequencing, but results of genetic testing were not available for this patient. There were no protein-truncating variants in other high-penetrance CRC susceptibility genes. Missense variants were identified, however, and those with low (< 5%) or unknown MAF were queried in the InSiGHT mutation database (`https://www.insight-group.org/variants/databases/`). Each missense variant was found in the InSiGHT database, but none of them had been classified as unambiguously pathogenic.

6.1.2 Prevalence of hereditary cancer syndromes

Sixteen of the 38 patients (42%; 95% CI, 26%-59%) had highly penetrant CRC susceptibility syndromes. The most frequent genetic disorder was LS, which was present in 12 of the 38 cases (32%). There were seven mutations in *MLH1* (7/38; 18%), four in *MSH2* (4/38; 11%) and one in *MSH6* (1/38; 2.6%). Early-onset MSI CRC was related to LS in 12 of 14 cases (86%). One patient (s49; age at diagnosis, 33 years) had both LS and MEN-1. One patient with MSI CRC (c543; age at diagnosis, 30 years) met clinical criteria for X-linked agammaglobulinemia (XLA), but WES did not reveal protein-coding variants in *BTK*, which is frequently mutated in XLA. XLA is a hereditary immunodeficiency syndrome associated with incomplete B-cell maturation and decreased immunoglobulin production (`https://www.omim.org/entry/300755`).

6.1.3 Clinical characteristics

By the time of diagnosis, most patients had developed either regional or distant metastases (Dukes' stage C or D, 61%, 23/38). The distribution of Dukes' stages was as follows: stage A, 6 cases (16%); stage B, 9 cases (24%); stage C, 17 cases (45%); and stage D, 6 cases (16%). No significant difference in Dukes' stage was found between syndromic and nonsyndromic cases ($p = 0.88$ by Wilcoxon's rank sum test). Due to the prevailing clinical practices at the time of sample collection, TNM stages were not available for most patients. Tumor grades had been recorded in the pathology reports of 35 of the 38 patients; 8.6% (3/35) of the tumors were well-differentiated (grade 1), 74% (26/35) were moderately differentiated (grade 2) and 17% (6/35) were poorly differentiated (grade 3). Seventy-four percent (28/38) of the primary tumors were distal to the splenic flexure. According to the pathology reports, none of the 38 patients had IBD. One patient (s124; age at diagnosis, 34 years) presented with two synchronous CRCs and one patient (c138; age at diagnosis, 39 years) had been diagnosed with metachronous CRC two years before sample collection; both s124 and c138 had LS.

6.1.4 Family history

Complete family histories of cancer in first-degree relatives were obtained shortly after surgery by linking data from official population registries and the Finnish Cancer Registry. Eleven patients (11/38, 29%) had familial CRC. Among those with familial early-onset CRC, 73% (8/11; 95% CI, 39%-94%) had a high-penetrance mutation, including seven patients with LS and one patient with FAP. Of those with negative family history, 30% had a hereditary CRC syndrome (8/27; 95% CI, 14%-50%). In contrast, only 14% (3/22) of those without well-defined predisposition syndromes had familial CRC, and none of them had synchronous or metachronous CRC. One patient with FAP (c231; age at diagnosis, 34) had two first-degree relatives with CRC, while the first-degree relatives of the other two patients with FAP did not have CRC.

6.2 Rare variants in early-onset colorectal cancer patients (II)

Twenty-two of the 38 early-onset CRC patients (58%) had not been diagnosed with known predisposing conditions. Because of their young age, we suspected that these patients may carry undiscovered protein-coding variants that influence CRC risk. It was also hypothesized that the genetic homogeneity of the Finnish population isolate may facilitate the identification rare disease-causing variants.

After quality control procedures, a total of 856,325 protein-coding variants were called in the 22 early-onset CRC patients. Of these, 203,699 (24%) were nonsynonymous and 6,120 (0.71%) were protein-truncating. Among the protein-truncating variants, the numbers (proportions) of nonsense, splice site and frameshift variants were 1,920 (31%), 1,613 (26%), and 2,587 (42%), respectively. Allele frequencies in 3,374 Finnish individuals and 58,112 non-Finnish individuals were obtained from the Exome Aggregation Consortium (ExAC) database (Cambridge, MA; `http://exac.broadinstitute.org`, accessed in November 2014). We hypothesized that early-onset CRC patients may carry rare variants that predispose to CRC in a dominant manner. Therefore, we searched for variants with MAF < 0.1% in both Finnish and non-Finnish controls. The possibility that early-onset CRC patients may have recessive predispositions was also considered. To this end, we examined rare homozygous genotypes that were not present in any of the 61,486 controls. To find further evidence of variant pathogenicity, we analyzed a validation set of 95 nonsyndromic familial CRC cases who were also of Finnish ancestry and had undergone WES in a previous study (Gylfe et al., 2013). The majority (85/95, 89%) of the familial CRC cases had one first-degree with CRC. Main results are summarized in Tables 6.1 and 6.2.

Table 6.1: Rare protein-coding variants in the 117 CRC patients with unknown etiology. *GRCh37, Genome Reference Consortium human genome build 37; YCRC, young CRC patients; FCRC, familial CRC patients.*

Gene	**Variant**	**GRCh37**	**YCRC**	**FCRC**	**ExAC**
ADAMTS4	c.1618delG	1:161163547	1/22	1/95	2/61,486 (0.0033%)
CYTL1	c.327+2T>A	4:5018561	1/22	1/95	19/61,486 (0.031%)
SYNE1	c.1941dupT	6:152784643	1/22	0/95	2/61,486 (0.0033%)
SYNE1	c.5568delC	6:152738004	0/22	1/95	0/61,486 (0%)
ACSL5	p.Pro71Leu	10:114154748	2/22	0/95	15/61,486 (0.024%)
INTS5	p.Pro922Leu	11:62414787	2/22	0/95	2/61,486 (0.0033%)
MCTP2	c.1488+1G>C	15:94911021	1/22	5/95	123/61,486 (0.20%)
ARHGAP12	p.Cys199Ser	10:32197188	1/22	0/95	30/61,486 (0.049%)
ATM	p.Ser333Phe	11:108117787	1/22	1/95	156/61,486 (0.25%)
DONSON	p.Glu471Lys	21:34951808	1/22	2/95	159/61,486 (0.26%)
ROS1	p.Ser370Pro	6:117715381	1/22	0/95	203/61,486 (0.33%)

6.2.1 Recurrent loss-of-function variants

In this series of 22 early-onset CRC cases, we did not find genes with recurrent LoF variants with MAF < 0.1%. Therefore, we examined rare LoF variants that were observed in one early-onset CRC patient and at least one of the 95 familial CRC patients. Three genes displayed rare LoF variants in both sample sets: *ADAMTS4* (c.1618delG), *CYTL1* (c.327+2T>A) and *SYNE1* (c.1941dupT and c.5568delC). Each of these three genes harbored rare LoF variants in one early-onset CRC case and one familial CRC case. The variants were validated by Sanger sequencing. The familial CRC patients with LoF variants in *ADAMTS4*, *CYTL1* and *SYNE1* had been diagnosed at 75, 86 and 84 years of age, respectively.

6.2.2 Recurrent missense variants

Next, we analyzed the 22 early-onset CRC patients for recurrent missense variants with MAF $< 0.1\%$. Two rare missense variants, each present in two of the 22 early-onset CRC patients, were identified: *ACSL5* p.Pro71Leu and *INTS5* p.Pro922Leu. Both variants were validated by Sanger sequencing. Neither of the variants was found in the validation set of 95 familial CRC cases. PolyPhen-2 classified *INTS5* p.Pro922Leu as possibly damaging and *ACSL5* p.Pro71Leu as benign, whereas SIFT classified both *INTS5* p.Pro922Leu and *ACSL5* p.Pro71Leu as deleterious. Neither of these missense variants were found in the validation set of 95 familial CRC patients.

6.2.3 Biallelic loss-of-function and missense variants

After quality control procedures, we observed a total of 69,503 nonsilent homozygous variants in the 22 early-onset CRC patients. Excluding variants with homozygous genotypes in the 61,486 controls, we identified four genes with rare homozygous missense variants: *ARHGAP12*, *ATM*, *DONSON* and *ROS1*. One early-onset CRC case had a homozygous splice site variant in *MCTP2* (c.1488+1G>C), which was validated by Sanger sequencing. All of these variants were rare in the Finnish population (MAF <1% in the 3,374 Finnish controls). None of the 95 familial CRC patients were homozygous for these variants. However, five of the 95 familial CRC cases (5.3%) were heterozygous for *MCTP2* c.1488+1G>C. Using similar exclusion criteria as for homozygous variants, we did not find genes with compound heterozygous LoF variants. Compound heterozygous missense variants could not be analyzed in a comprehensive manner due to the lack of haplotype-resolved data.

6.2.4 Allelic imbalance

We hypothesized that some of the identified LoF or missense variants may display somatic LOH similar to classical TSGs. Paired cancer samples from each patient were analyzed for allelic imbalance by Sanger sequencing, which did not suggest allelic imbalance for any of the variants.

Table 6.2: Clinical and molecular characteristics of the 38 early-onset CRC patients. Mutation 1 is a 3.5 kb deletion comprising exon 16 of *MLH1* (Nyström-Lahti et al., 1995). *BMPR1A* g.391C>G has been shown to result in skipping of exon 1 of *BMPR1A* (Zhou et al., 2001). *Age, age at diagnosis (years); FH, family history of CRC.*

ID	**Age**	**Dukes**	**MSI**	**FH**	**Syndrome/gene(s)**	**Variant(s)**
c79	23	C	-	-	*BMPR1A*	g.391C>G
s30	27	A	+	-	*MLH1*	c.1975C>T
s17	27	C	+	+	*MLH1*	c.677+1G>T
s124	34	C	+	+	*MLH1*	Mutation 1
s171	35	B	+	-	*MLH1*	c.454-1G>A
c430	36	B	+	+	*MLH1*	Mutation 1
s1108	38	C	+	+	*MLH1*	Mutation 1
c615	39	B	+	+	*MLH1*	c.454-1G>A
s49	33	C	+	-	*MSH2*; MEN-1	Deletion of exons 1-7 (*MSH2*)
c54	35	D	+	-	*MSH2*	c.1387-1G>T
s108	38	A	+	+	*MSH2*	Deletion of exons 7-8
c138	39	D	+	-	*MSH2*	c.1807G>A
s95	31	C	+	+	*MSH6*	c.3013C>T
c520	29	C	-	-	FAP (clinical dx)	-
c231	34	A	-	+	*APC*	c.739C>T
s200	37	C	-	-	FAP (clinical dx)	-
s907	21	B	-	-	-	-
s1152	28	D	-	-	*SYNE1*	c.1941dupT
c386	30	C	-	+	-	-
c592	30	A	-	-	-	-
c543	30	B	+	-	XLA	-
c1066	31	C	-	-	*ACSL5*;*MCTP2*;*ROS1*	p.Pro71Leu; c.1488+1G>C; p.Ser370Pro
s1137	31	D	-	-	*CYTL1*;*INTS5*	c.327+2T>A; p.Pro922Leu
c768	32	C	+	-	*ACSL5*;*ATM*	p.Pro71Leu; p.Ser333Phe
s160	33	C	-	-	-	-
c907	34	A	-	-	*ADAMTS4*	c.1618delG
s1167	35	B	-	-	-	-
c206	36	C	-	+	-	-
c414	36	C	-	-	-	-
c690	36	B	-	-	-	-
s1151	36	D	-	-	*INTS5*	p.Pro922Leu
s281	37	B	-	-	*ARHGAP12*	p.Cys199Ser
s1165	37	C	-	-	-	-
c938	38	C	-	-	-	-
c1055	38	C	-	-	-	-
s154	38	A	-	-	-	-
c270	39	D	-	+	*DONSON*	p.Glu471Lys
c837	39	B	-	-	-	-

7 Genome-wide association study and meta-analysis of colorectal cancer

7.1 Multi-stage genome-wide association study (III)

7.1.1 Stage 1

In stage 1, 9,068,015 single-nucleotide variants were imputed and tested for association with CRC in the FIN series. The genomic inflation factor was 1.12, and genomic control was applied accordingly. rs73121704 at 12q14.3 (MAF, 0.860%) displayed the smallest p-value in stage 1 ($p = 4.07 \cdot 10^{-9}$). p-values for all other SNPs in stage 1 were above the genome-wide significance threshold of $5 \cdot 10^{-8}$. A Manhattan plot is shown in Figure 7.1.

7.1.2 Stage 2

Most promising SNPs from stage 1 were genotyped in five Nordic case-control series (STHLM2, Gothenburg, HUNT, Estonia, and FINRISK) comprising a total of 4,070 cases and 2,377 controls. Genotyping assays were designed for 40 variants from 20 loci - two variants per locus. rs992157 (2q35) was also genotyped in stage 2 because it had been recently reported as a CRC risk SNP (Orlando et al., 2016). Eleven of the 41 variants could not be genotyped because of difficult sequence context (seven variants) or assay failure (four variants). Therefore, 30 variants from 20 loci were genotyped successfully. The allele count of 6:73457627G>C was low in all five Nordic series (< 10). Pearson correlation coefficient (r^2) between IMPUTE2 genotype dosage and MassARRAY genotype was used to evaluate the accuracy of genotype imputation in 1,038 individuals from the FIN series. r^2 values for the 30 variants ranged from 0.816 to 1.00 with a median of 0.978.

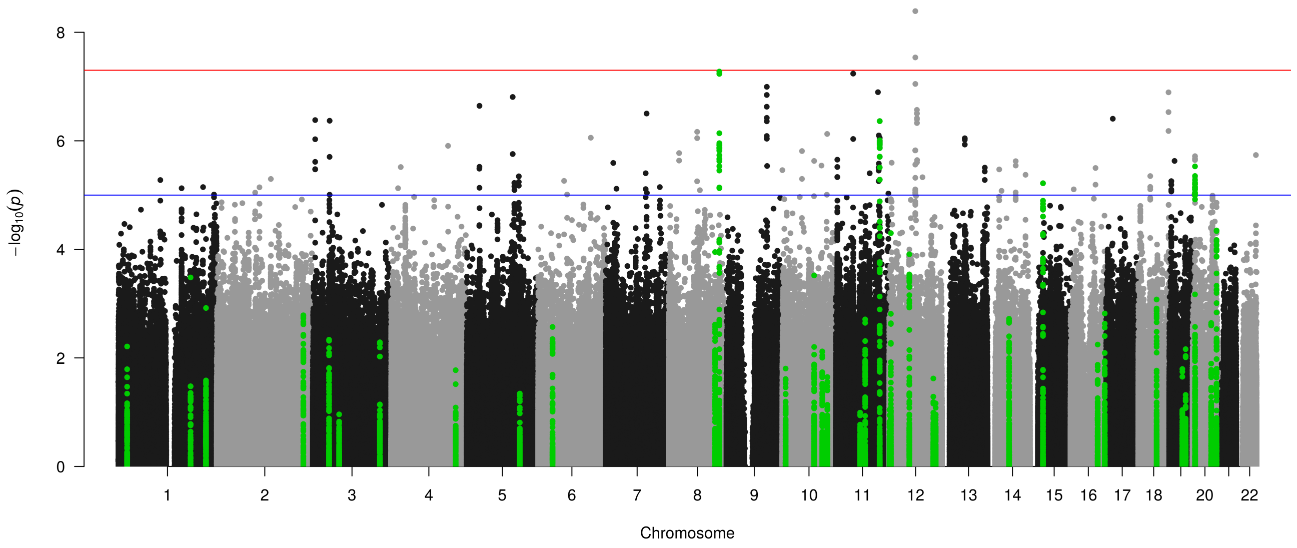


Figure 7.1: Manhattan plot for the FIN GWAS (stage 1 of study III). Thirty-eight previously published CRC susceptibility loci (see text) are highlighted in green.

7.1.3 Stage 3

To increase statistical power, we extracted summary statistics from previously published GWASs (Al-Tassan et al., 2015). These studies comprised 7,577 CRC cases and 9,979 controls of European ancestry (Al-Tassan et al., 2015). Data were available for 27 of the 30 variants that had been successfully genotyped in stage 2. The SNPs with missing data were rs150509351, rs186867472 and 6:73457627G>C.

7.1.4 Combined analysis

Combined analysis of stages 1-3 comprised 13,348 cases and 26,438 controls. Because of possible study heterogeneity, the primary meta-analysis was based on a random-effects model (Evangelou et al., 2013). Standard errors from each study were corrected according to the respective inflation factors (FIN, 1.11; COIN, 1.10; UK1, 1.03; Scotland1, 1.04; VQ58, 1.04; CCFR1, 1.03; and CCFR2, 1.08). Three SNPs from two loci showed genome-wide significant associations with CRC: rs10505477 (8q24.21, $p = 7.63 \cdot 10^{-14}$), rs6983267 (8q24.21, $p = 7.45 \cdot 10^{-13}$) and rs992157 (2q35, $p = 1.50 \cdot 10^{-9}$). One SNP, rs6589219 at 11q23.1, showed a suggestive association with CRC ($p = 9.14 \cdot 10^{-6}$). Each of these four SNPs represented previously published loci. The results of combined analysis are summarized in Table 7.1.

Table 7.1: Results of meta-analysis of 39,786 European-ancestry individuals.

Chr.	**SNP**	**Gene**	p**-value**	p_{het}	I^2	**Reference**
8q24.21	rs10505477-A	*CASC8*	$7.63 \cdot 10^{-14}$	0.144	34.4%	Tomlinson et al., 2007
8q24.21	rs6983267-G	*CASC8*	$7.45 \cdot 10^{-13}$	0.0985	37.7%	Zanke et al., 2007
2q35	rs992157-A	*PNKD*	$1.50 \cdot 10^{-9}$	0.777	0%	Orlando et al., 2016
11q23.1	rs6589219-G	*COLCA1*	$9.14 \cdot 10^{-6}$	0.153	36.5%	Tenesa et al., 2008

7.2 Replication of published colorectal cancer risk SNPs (III)

7.2.1 Replication of 38 colorectal cancer risk SNPs

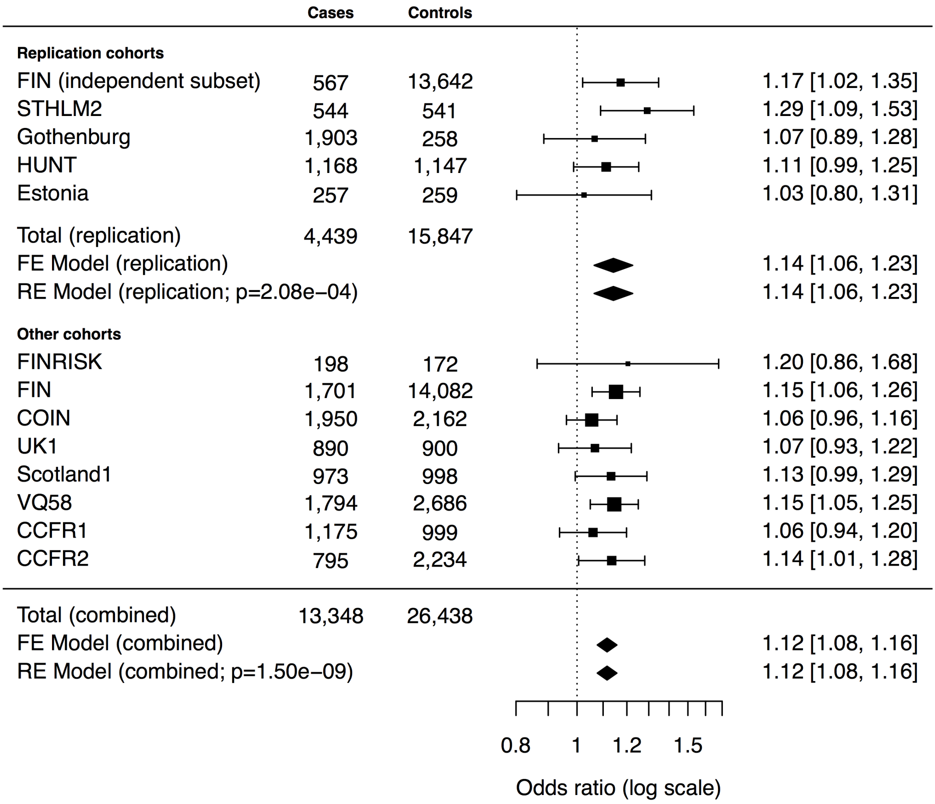
In stage 1, there was suggestive evidence of association ($p < 1 \cdot 10^{-5}$) for the known CRC risk SNPs rs10505477 ($p = 5.29 \cdot 10^{-8}$), rs6983267 ($p = 1.38 \cdot 10^{-6}$) and rs6589219 ($p = 4.34 \cdot 10^{-7}$). rs6589219 was considered a known CRC risk SNP because it is in strong LD with rs3802842 (Tenesa et al., 2008; r^2, 0.942 in 1,000 Genomes Phase 3 European populations). To replicate published disease associations more systematically, we analyzed a set of 38 SNPs that have been previously reported as CRC risk SNPs in European populations (Frampton et al., 2016; Orlando et al., 2016). Fourteen of the 38 SNPs (37%) were associated with CRC in the Finnish population with false-discovery rate (q) < 10%, and all of these 14 SNPs showed the same direction of effect as previously reported. The results are summarized in Table 7.2.

Table 7.2: Replication of published CRC risk SNPs in the FIN series.

Chr.	**SNP**	**Gene**	**q-value**	**Reference**
11q23.1	rs3802842-C	*COLCA1*	$1.77 \cdot 10^{-5}$	Tenesa et al., 2008
8q24.21	rs6983267-G	*CASC8*	$1.77 \cdot 10^{-5}$	Tomlinson et al., 2007
8q24.21	rs7014346-A	*CASC8*	$1.77 \cdot 10^{-5}$	Tenesa et al., 2008
20p12.3	rs961253-A	*CASC20*	$6.92 \cdot 10^{-5}$	COGENT Study, 2008
15q13.3	rs4779584-T	*SCG5*	$1.29 \cdot 10^{-3}$	Jaeger et al., 2008
10q22.3	rs704017-G	*ZMIZ1-AS1*	$1.91 \cdot 10^{-3}$	Zhang et al., 2014b
18q21.1	rs7229639-A	*SMAD7*	$7.96 \cdot 10^{-3}$	Zhang et al., 2014a
2q35	rs992157-A	*PNKD*	$7.96 \cdot 10^{-3}$	Orlando et al., 2016
8q23.3	rs16892766-C	*EIF3H*	0.0113	Tomlinson et al., 2008
14q22.2	rs4444235-C	*BMP4*	0.0231	COGENT Study, 2008
6p21.2	rs1321311-A	*PANDAR*	0.0231	Dunlop et al., 2012
20q13.33	rs4925386-C	*LAMA5*	0.0501	Houlston et al., 2010
10q24.2	rs1035209-T	*SLC25A28*	0.0536	Whiffin et al., 2014
11q13.4	rs3824999-G	*POLD3*	0.0604	Dunlop et al., 2012

7.2.2 Independent replication of rs992157 (2q35)

Because a subset of the FIN series had contributed to the original discovery of the association between rs992157 (2q35) and CRC (Orlando et al., 2016), rs992157 was independently replicated in 4,439 CRC cases and 15,847 controls from five Nordic case-control series (STHLM2, Gothenburg, HUNT, Estonia and part of the FIN series) that had not been previously studied for this association. Genomic control was applied in the FIN series (inflation factor, 1.11), but inflation factors could not be estimated in other Nordic series because only 30 SNPs had been genotyped. Regression coefficients from logistic regression models were combined with the use of random-effects meta-analysis. There was no notable study heterogeneity ($p_{het} = 0.462$, $I^2 = 0\%$). After Bonferroni correction for the 30 variants that were genotyped in the MassARRAY experiment ($\alpha = 0.05/30 \approx 0.00167$), rs992157 was significantly associated with CRC with an OR of 1.14 ($p = 2.08 \cdot 10^{-4}$; 95% CI, 1.06-1.23). Consistent with prior results, the alternative allele (A) conferred a higher risk of CRC than the reference allele (G). There was near-perfect correlation between MassARRAY genotype and IMPUTE2 genotype dosage in the FIN series (r^2, 1.00). A Forest plot for rs992157 is shown in Figure 7.2.



	Cases	Controls	
Replication cohorts			
FIN (independent subset)	567	13,642	1.17 [1.02, 1.35]
STHLM2	544	541	1.29 [1.09, 1.53]
Gothenburg	1,903	258	1.07 [0.89, 1.28]
HUNT	1,168	1,147	1.11 [0.99, 1.25]
Estonia	257	259	1.03 [0.80, 1.31]
Total (replication)	4,439	15,847	
FE Model (replication)			1.14 [1.06, 1.23]
RE Model (replication; p=2.08e−04)			1.14 [1.06, 1.23]
Other cohorts			
FINRISK	198	172	1.20 [0.86, 1.68]
FIN	1,701	14,082	1.15 [1.06, 1.26]
COIN	1,950	2,162	1.06 [0.96, 1.16]
UK1	890	900	1.07 [0.93, 1.22]
Scotland1	973	998	1.13 [0.99, 1.29]
VQ58	1,794	2,686	1.15 [1.05, 1.25]
CCFR1	1,175	999	1.06 [0.94, 1.20]
CCFR2	795	2,234	1.14 [1.01, 1.28]
Total (combined)	13,348	26,438	
FE Model (combined)			1.12 [1.08, 1.16]
RE Model (combined; p=1.50e−09)			1.12 [1.08, 1.16]

Figure 7.2: Forest plot for rs992157 (2q35).

Discussion

8 Genetic background of early-onset colorectal cancer (I, II)

CRC is the fifth most common cancer type in young adults aged 20 to 39 years after breast cancer, cervical cancer, thyroid cancer and leukemias (Fidler et al., 2017). Although the overall incidence of CRC has declined over the last decades, CRC and especially rectal cancer have become more common in young adults (Siegel et al., 2017). The causes of these epidemiologic changes are unknown, but they may be related to modifiable risk factors such as diet, physical inactivity, obesity and lack of screening for CRC in young adults (Siegel et al., 2009; Bailey et al., 2015).

The rising incidence of CRC in young adults has drawn interest into genetic risk assessment, which could help to identify individuals who may benefit from earlier screening (The Lancet Oncology, 2017;18(4):413). The prevalence of hereditary syndromes in early-onset CRC cases is high; it has been estimated at 34.7% (95% CI, 28.1%-41.9%) in those younger than 35 years at diagnosis and at 16.0% (95% CI, 12.8%-19.8%) in those younger than 50 years at diagnosis, with LS being the most frequent genetic disorder (Mork et al., 2015; Pearlman et al., 2017). In contrast, less than 5% of unselected CRC patients have underlying high-penetrance syndromes (Aaltonen et al., 2007). GI polyposis syndromes are usually diagnosed on the basis of clinical and histopathologic criteria, whereas tumor MSI identifies patients with possible LS. Hereditary syndromes may, however, present atypically and may lack characteristic clinicopathologic features (Sweet et al., 2005). In these cases, syndrome identification may depend on direct genetic testing. In recent years, diagnostic NGS has become widely available and may help uncover pathogenic germline variants in patients without clear clinical features of hereditary cancer syndromes.

We studied an unselected series of 1,514 Finnish CRC patients, 38 of whom had been diagnosed before age 40 years. Twenty-three patients with unexplained etiology were analyzed for nonsynonymous variants in 10 high-penetrance CRC predisposition genes (*MLH1*, *MSH2*, *MSH6*, *PMS2*, *APC*, *MUTYH*, *SMAD4*, *BMPR1A*, *STK11* and *PTEN*). WES revealed only one additional pathogenic variant in *MLH1*. Therefore, WES provided little added value in diagnosing the underlying conditions as compared with standard diagnostic methods. WES did, however, identify missense VUSs. VUSs may complicate genetic counseling and clinical management but can be reclassified when sufficient evidence accumulates. WES has limited sensitivity for structural variants, which can be detected more reliably with MLPA, WGS or NGS panels. WES may be more cost-efficient than WGS, and the diagnostic yield of NGS panels depends on which genes are targeted (Sun et al., 2015).

Consistent with prior literature, the prevalence of hereditary CRC syndromes was high (42%; 95% CI, 26%-59%). The most prevalent syndrome was LS, which was diagnosed in 12 of the 38 patients (32%). Correspondingly, the proportion of MSI CRCs was high (37%, 14/38), and LS was present in 86% (12/14) of patients with early-onset MSI CRC. Therefore, early-onset MSI CRC should raise strong suspicion of LS. FAP had been diagnosed in three patients (7.9%) and JP in one patient (2.6%), reflecting the lower incidence rates of these syndromes. These estimates may be influenced by referral bias because 31% of the patients (472/1,514) were not part of the population-based sample collection (May 1994 - June 1998).

By the time of diagnosis, most patients (61%) had developed either local or distant metastases, but there was no notable difference in tumor stage between syndromic and nonsyndromic patients. Previous studies support an association between early age at diagnosis and advanced stage, which may be due to diagnostic delay or aggressive tumor biology (O'Connell et al., 2004a; Taggarshe et al., 2013). Both provider- and patient-related factors may delay cancer diagnoses in young adults (Bleyer et al., 2006). Early age of onset does not seem to be independently associated with poor CRC-specific outcomes, however (O'Connell et al., 2004b; Hubbard et al., 2012). Poor differentiation and mucinous histology, both of which are associated with aggressive clinical course, are more common among early-onset CRC patients (Griffin et al., 1991).

Official population registries and the Finnish Cancer Registry were used to obtain accurate data on family history. In this series, familial early-onset CRC was associated with a high proportion of hereditary cancer syndromes (73%; 8/11), but these syndromes were also common in patients with negative family history (30%; 8/27). Registry-based family histories are generally less liable to false-positive and false-negative reports than self-reported family histories (Murff et al., 2004).

One patient with MSI CRC (c543; age at diagnosis, 30 years) met clinical criteria for XLA. It has been previously proposed that XLA is associated with an increased risk of CRC (Meer et al., 1993). MSI gives rise to a large number of neoantigens, and an intact immune system may be particularly important for the elimination of MSI CRCs or their precursors. This hypothesis is consistent with MSI being a strong predictive marker for benefit from cancer immunotherapy (Le et al., 2017). Another patient (s49; age at diagnosis, 33 years) had both LS and MEN-1. This patient had been diagnosed with CRC, pituitary adenoma, hyperparathyroidism and breast hypertrophy, but there was no clear evidence of interaction between these two gene defects.

Among the 22 early-onset CRC patients, there were no genes with recurrent LoF variants with MAF $< 0.1\%$. The lack of such genes, together with negative family history in the majority of the patients (86%) suggests that the genetic etiology of early-onset CRC in the absence of known hereditary cancer syndromes is relatively complex. Possible etiologies include rare variants with intermediate penetrance, *de novo* mutations, recessive inheritance, polygenic inheritance and environmental risk factors. Part of the risk is naturally explained by replicative mutations and other stochastic events (Tomasetti et al., 2015a). We also considered genes with LoF variants in one early-onset CRC patient and at least one familial CRC patient. Three genes - *ADAMTS4*, *CYTL1* and *SYNE1* - displayed rare LoF variants in one early-onset CRC patient and one familial CRC patient. All three familial CRC patients with LoF variants in these genes had late-onset CRC (age at diagnosis $\geq$ 75 years). The known functions of *SYNE1*, *CYTL1* and *ADAMTS4* are related to intracellular spatial organization, CD34+ mononuclear cells and cartilage homeostasis, respectively (Tortorella et al., 2000; Zhang et al., 2001; Liu et al., 2000).

Lek et al., 2016, derived the probability of LoF intolerance (pLI) score by comparing observed and expected numbers of protein-truncating variants in human genes. Genes with pLI score > 0.9 were classified as LoF intolerant and those with pLI score < 0.1 as LoF tolerant. pLI scores for *ADAMTS4*, *CYTL1* and *SYNE1* were 0.01, 0.00 and 0.00, respectively. Although many high-penetrance cancer genes have high pLI scores (e.g., pLI scores for *APC* and *MLH1* are 1.00 and 0.74, respectively), some cancer susceptibility genes are LoF tolerant (e.g., both *BRCA1* and *BRCA2* have a pLI score of 0.00).

Most genes identified in this study (*ADAMTS4*, *CYTL1*, *SYNE1*, *ACSL5*, *INTS5*, *MCTP2*, *ARHGAP12* and *DONSON*) have not been previously implicated in human cancer as defined by the Cancer Gene Census (`http://cancer.sanger.ac.uk/cancergenome/projects/census/`, accessed June 2018). There are a number of plausible explanations. First, because of the exploratory nature of the study, we cannot firmly conclude whether the variants are pathogenic or not. Random chance may have played a role. Second, previous linkage studies and GWASs have not been ideal for identifying rare moderate-penetrance variants, and adequately powered NGS studies are only beginning to emerge. Therefore, the genes shortlisted here are early candidates for incompletely penetrant CRC susceptibility. Third, some of the variants may be relatively population-specific, making them difficult to discover in non-Finnish populations. Two of the identified genes, *ATM* and *ROS1*, do have established roles in human cancer. Biallelic LoF mutations in *ATM*, a DNA damage response kinase, are associated with autosomal recessive ataxia telangiectasia (Savitsky et al., 1995). Candidate gene studies and GWASs suggest that monoallelic carriers of *ATM* mutations are at increased risk of breast, gastric, pancreatic and prostate cancers (Thompson et al., 2005; Renwick et al., 2006; Helgason et al., 2015), but evidence of association with CRC is insufficient (Thompson et al., 2005). *ROS1* is a tyrosine kinase and a proto-oncogene, which makes it a somewhat unlikely candidate gene for cancer predisposition. *ACSL5* has been shown to inhibit Wnt signaling in intestinal surface epithelia through palmitoylation of the Wnt2B protein (Klaus et al., 2014). Therefore, impaired function of *ACSL5* may have proliferative and antiapoptotic effects. *INTS5* encodes a subunit of the integrator complex, which regulates RNA polymerase II -mediated transcription. The integrator complex has multiple functions - especially RNA processing and genome maintenance - that may be relevant for cancer (Federico et al., 2017). One early-onset CRC patient had a biallelic splice site variant in *MCTP2* (c.1488+1G>C), and five familial CRC patients (5/95, 5.3%) were heterozygous for *MCTP2* c.1488+1G>C. Complete gene inactivation is rare in human genomes, and the reported function of *MCTP2* in cardiac development questions the nature of c.1488+1G>C as a LoF variant (Lalani et al., 2013; Sulem et al., 2015). However, *MCTP2* appears to be LoF tolerant (pLI, 0.00).

In conclusion, syndromic early-onset CRC was characterized by MSI, GI polyposis and positive family history. The etiology of nonsyndromic early-onset CRC remains poorly understood - these patients were characterized by MSS CRC, lack of GI polyposis and negative family history. WES has become widely available in clinical practice, but here it provided little additional clues for genetic diagnosis as compared with the current standard strategies of MSI testing and workup for GI polyposis. WES of the 22 unexplained early-onset CRC cases did not suggest any single genetic defect that would explain a major part of cancer risk in this patient group, but the identified rare variants may be of interest in future genetic studies on multifactorial CRC susceptibility. Larger studies are needed to further define the diagnostic value of WES and to elucidate the contribution of rare variants to early-onset CRC.

9 Common variants in colorectal cancer predisposition (III)

Beginning in 2007, GWASs have provided convincing evidence that common variants explain part of the variation in CRC risk, thereby supporting the common disease-common variant hypothesis (Reich et al., 2001). Each common variant has a modest additive or multiplicative effect on disease risk, and most individuals in the general population carry multiple alleles with opposing effects (Holst et al., 2010). Common variants have been estimated to explain between 7.4% and 19% of the heritability of CRC, but currently known risk SNPs explain only around 10% of the heritability that can be ascribed to common variants (Jiao et al., 2014; Frampton et al., 2016). Therefore, efforts to identify additional CRC risk SNPs remain relevant. Furthermore, GWASs have identified disease associations in unexpected regions of the genome, overcoming the limitations of candidate gene studies (Houlston et al., 2001). For example, few genes involved in DNA repair have been identified by GWASs, although they are mutated in multiple hereditary CRC syndromes. Finally, studies in isolated populations may be of interest (Kristiansson et al., 2008) because most GWASs of CRC have been conducted in outbred populations.

We conducted a GWAS of CRC in the Finnish population (stage 1), replicated results in other European-ancestry populations (stages 2-3) and meta-analyzed a total of 13,348 cases and 26,438 controls. The recently identified CRC risk SNP rs992157 (2q35 near *PNKD* and *TMBIM1*) and 37 other SNPs that have been associated with CRC in European populations were analyzed in samples that were independent of previous studies.

The association between rs992157 (2q35) and CRC was replicated in five independent Nordic datasets (STHLM2, Gothenburg, HUNT, Estonia and an independent subset of the FIN series) that had not been previously studied for this association ($p = 2.08 \cdot 10^{-4}$; OR, 1.14; 95% CI, 1.06-1.23). The effect size was consistent with that reported by Orlando et al., 2016, and the association was genome-wide significant in combined analysis ($p = 1.50 \cdot 10^{-9}$; OR, 1.12; 95% CI, 1.08-1.16). The risk allele for CRC, rs992157-A, is correlated with rs2382817-C, which confers protection from IBD (r^2, 0.826 in 1,000 Genomes Phase 3 European populations; Jostins et al., 2012). Considering that IBD is a risk factor for CRC, these observations suggest genetic pleiotropy and point to a complex relationship between inflammatory pathways and cancer. Interestingly, rs992157-A is also associated with increased adult human height (Wood et al., 2014).

In addition to 2q35, twelve other loci were associated with CRC in the FIN series (6p21.2, 8q23.3, 8q24.21, 10q22.3, 10q24.2, 11q13.4, 11q23.1, 14q22.2, 15q13.3, 18q21.1, 20p12.3 and 20q13.33; for the original discovery studies, see Table 3.2), which suggests that the genetic architecture of complex CRC susceptibility is similar between the Finnish population isolate and outbred populations. Replication of findings from GWASs is important because it reduces the risk that proposed disease associations are due to random error or uncontrolled bias (Kraft et al., 2009). The lack of novel genome-wide significant CRC risk SNPs in this study may reflect the small effects and/or low allele frequencies of the variants that remain undiscovered.

This study has several limitations. Samples from the Finnish CRC Collection and the FINRISK Study were genotyped with different SNP arrays, which is a potential source of bias. rs73121704 at 12q14.3 (MAF, 0.860%) was strongly associated with CRC in the FIN series ($p = 4.07 \cdot 10^{-9}$) but not in other studies (random-effects $p = 0.466$, fixed-effect $p = 0.122$), which raises concern on imputation accuracy. Of note, 63% (24/38) of the previously published CRC SNPs failed to replicate in this study. A main reason for non-replication may have been insufficient statistical power (`http://csg.sph.umich.edu/abecasis/cats/gas_power_calculator/`), but also possible are genotyping error, different LD patterns across populations and spurious associations (Kraft et al., 2009). Because of financial constraints, a relatively small number of variants were studied in stages 2 and 3.

In conclusion, we replicated the associations of 14 previously published SNPs with CRC in the Finnish population, but novel CRC risk SNPs were not identified. These results validate findings from previous studies and inform the design of future GWASs in isolated populations.

Concluding remarks and future prospects

The genetic basis of CRC is an extensively studied topic. Even so, the majority of the heritability of CRC remains unexplained. Upcoming GWASs with larger sample sizes are likely to reveal a long tail of additional CRC risk loci that each explain smaller and smaller proportions of the total heritability. On the other hand, new hereditary CRC syndromes may be found in families with very rare genetic defects. As the cost of genome sequencing continues to decline, NGS-based association studies will clarify the contribution of moderate-penetrance variants to CRC susceptibility. The expanding use of clinical NGS in cancer patients will produce vast amounts of genomic data as a byproduct of clinical practice, which will provide a valuable resource for research in cancer genomics. Over time, the focus of research may shift towards understanding the biological mechanisms of genetic associations, as well as developing clinical applications.

The functional characterization of GWAS loci is a challenging but important goal. Despite the modest effect sizes of common SNPs, the underlying biological mechanisms may be conceptually important for developing pharmacological interventions with clinically significant effects (Nelson et al., 2015). Even in the absence of detailed biological knowledge, genetic associations offer direct opportunities for clinical applications. Although risk-reducing interventions have been successful in patients with hereditary cancer syndromes, it is not known if PRSs alone will be clinically useful. Instead, polygenic effects can be incorporated into more comprehensive risk models that may also include family history and environmental and behavioral risk factors (Torkamani et al., 2018). Such statistical models may allow the identification of individuals whose multifactorial risk of CRC is sufficiently high to warrant clinical intervention.

Acknowledgments

This work was carried out at the Department of Medical and Clinical Genetics and the Research Programs Unit, Faculty of Medicine, University of Helsinki, during 2012-2018. I wish to thank the present and former heads of the Department of Medical and Clinical Genetics and the Research Programs Unit for providing excellent research facilities.

I am sincerely grateful to professor Lauri Aaltonen for supervising this thesis and for giving me the chance to work in his outstanding research group. Lauri's unique expertise, curiosity and integrity made it a privilege to be one of his students. I want to thank my second supervisor, Sari Tuupanen, for providing excellent practical guidance, advice and thoughtful comments on manuscripts, and my third supervisor, Kimmo Palin, for his great problem-solving skills, computational support and encouragement to finish this work. I appreciate the time and effort my thesis reviewers, Asta Försti and Kati Kristiansson, put into assessing this work.

All collaborators and co-authors who contributed to this thesis are warmly thanked: Jukka-Pekka Mecklin, Heikki Järvinen, Anna Lepistö, Laura Renkonen-Sinisalo, Ari Ristimäki, Jan Böhm, Minna Taipale, Kaisa Silander, Aarno Palotie, Samuli Ripatti, Eero Pukkala, Veikko Salomaa, Pekka Jousilahti, Antti-Pekka Sarin, Elinor Bexe Lindskog, Yvonne Wettergren, Kristian Hveem, Eivind Ness-Jensen, Andres Metspalu, Neeme Tõnisson, Richard S. Houlston, Ian P. Tomlinson, Malcolm G. Dunlop, Philip J. Law, Aung K. Win, Nada A. Al-Tassan, Ella Barclay, Daniel D. Buchanan, Harry Campbell, Graham Casey, Jeremy P. Cheadle, David Conti, Susan M. Farrington, Steven Gallinger, John Hopper, Shelley Idziaszczyk, Mark A. Jenkins, Richard Kaplan, David Kerr, Rachel Kerr, Lynn Martin, Tim S. Maughan, Brian F. Meyer, Polly A. Newcomb, Giulia Orlando, Claire Palles, Fredrick R. Schumacher, Christopher G. Smith, Albert Tenesa, Maria N. Timofeeva and Salma M. Wakil.

It was a great pleasure to work in the "CRC room" together with Alexandra Gylfe, Johanna Kondelin, Ulrika Hänninen, Tatiana Cajuso, Jaana Tolvanen, Aurora Taira and Linda van den Berg. Being involved in so many research projects besides my thesis work was fascinating.

Postdocs - Niko Välimäki, Esa Pitkänen, Auli Karhu, Pia Vahteristo, Outi Kilpivaara, Mervi Aavikko, Netta Mäkinen, Linda Forsström, Janne Ravantti, Simona Bramante and Davide Berta - are warmly thanked for sharing their expertise and generous support.

My appreciation goes to Sini Marttinen, Sirpa Soisalo, Marjo Rajalaakso, Heikki Metsola, Inga-Lill Åberg, Iina Vuoristo, Alison Ollikainen and Mairi Kuris for their excellent technical support.

Conferences and lab visits were inspiring experiences. I want to thank Johanna Kondelin and Hanna Heinonen for our visit to Baltimore/Boston in 2015 and Tatiana Cajuso, Ulrika Hänninen, Jaana Tolvanen, Netta Mäkinen and Tuomas Heikkinen for our visit to Washington, D.C., in 2017.

I want to thank present and former members of the Aaltonen group - Riku Katainen, Iikki Donner, Lauri Sipilä, Eevi Kaasinen, Miika Mehine, Jiri Hamberg, Päivi Sulo, Roosa-Maria Plaketti, Heli Rauanheimo, Elina Heliövaara, Silva Saarinen, Heli Lehtonen, Iina Niittymäki, Kati Kämpjärvi, Javier Núñez-Fontarnau, Swee Wong and Heikki Ristolainen - for creating a positive and cheerful work atmosphere. Many thanks to Riku Katainen for creating BasePlayer.

I would like to thank the Biomedicum Helsinki Foundation, The Finnish Medical Foundation and Cancer Society of Finland for personal grants.

I would like to express my gratitude to the patients, families and healthcare providers who made this study possible.

My deepest thanks go to my family and friends for their unconditional support and encouragement throughout the project.

Helsinki, December 2018

Tomas Tanskanen

Bibliography

Aaltonen, L. A. et al. (1993). “Clues to the pathogenesis of familial colorectal cancer.” In: *Science* 260.5109, pp. 812–816.

Aaltonen, L. A. et al. (1998). “Incidence of hereditary nonpolyposis colorectal cancer and the feasibility of molecular screening for the disease.” In: *The New England journal of medicine* 338.21, pp. 1481–1487.

Aaltonen, L. et al. (2007). “Explaining the familial colorectal cancer risk associated with mismatch repair (MMR)-deficient and MMR-stable tumors.” In: *Clinical cancer research: an official journal of the american association for cancer research* 13.1, pp. 356–361.

Abbas, T. and A. Dutta (2009). “p21 in cancer: intricate networks and multiple activities”. In: *Nature reviews. Cancer* 9.6, pp. 400–414.

Aird, D. et al. (2011). “Analyzing and minimizing PCR amplification bias in Illumina sequencing libraries.” In: *Genome biology* 12.2, R18.

Al-Tassan, N. et al. (2002). “Inherited variants of MYH associated with somatic G:C–¿T:A mutations in colorectal tumors.” In: *Nature genetics* 30.2, pp. 227–232.

Al-Tassan, N. A. et al. (2015). “A new GWAS and meta-analysis with 1000Genomes imputation identifies novel risk variants for colorectal cancer.” In: *Scientific reports* 5, p. 10442.

Alexandrov, L. B. et al. (2013). “Signatures of mutational processes in human cancer.” In: *Nature* 500.7463, pp. 415–421.

Amin, M. B. et al. (2016). *AJCC Cancer Staging Manual.* Springer.

Amos, C. I. et al. (1997). “Fine mapping of a genetic locus for Peutz-Jeghers syndrome on chromosome 19p.” In: *Cancer research* 57.17, pp. 3653–3656.

Armitage, P. and R. Doll (1954). “The age distribution of cancer and a multistage theory of carcinogenesis.” In: *British journal of cancer* 8.1, pp. 1–12.

Aune, D. et al. (2011a). “Dietary fibre, whole grains, and risk of colorectal cancer: systematic review and dose-response meta-analysis of prospective studies.” In: *BMJ (Clinical research ed.)* 343.nov10 1, pp. d6617–d6617.

Aune, D. et al. (2011b). “Nonlinear reduction in risk for colorectal cancer by fruit and vegetable intake based on meta-analysis of prospective studies”. In: *Gastroenterology* 141.1, pp. 106–118.

Bailey, C. E. et al. (2015). “Increasing disparities in the age-related incidences of colon and rectal cancers in the United States, 1975-2010.” In: *JAMA surgery* 150.1, pp. 17–22.

Balding, D. J., M. Bishop, and C. Cannings (2007). *Handbook of Statistical Genetics.* Wiley-Interscience.

Beaugerie, L. and S. H. Itzkowitz (2015). "Cancers complicating inflammatory bowel disease." In: *The New England journal of medicine* 372.15, pp. 1441–1452.

Bellido, F. et al. (2016). "POLE and POLD1 mutations in 529 kindred with familial colorectal cancer and/or polyposis: Review of reported cases and recommendations for genetic testing and surveillance". In: *Genetics in medicine : official journal of the American College of Medical Genetics* 18.4, pp. 325–332.

Bennett, S. T. et al. (2005). "Toward the 1,000 dollars human genome." In: *Pharmacogenomics* 6.4, pp. 373–382.

Bibbins-Domingo, K. and U.S. Preventive Services Task Force (2016). *Aspirin Use for the Primary Prevention of Cardiovascular Disease and Colorectal Cancer: U.S. Preventive Services Task Force Recommendation Statement.* American College of Physicians.

Bisgaard, M. L. et al. (1994). "Familial adenomatous polyposis (FAP): frequency, penetrance, and mutation rate." In: *Human Mutation* 3.2, pp. 121–125.

Björk, J et al. (1999). "Epidemiology of familial adenomatous polyposis in Sweden: changes over time and differences in phenotype between males and females." In: *Scandinavian journal of gastroenterology* 34.12, pp. 1230–1235.

Bleyer, A., T. Budd, and M. Montello (2006). "Adolescents and young adults with cancer: the scope of the problem and criticality of clinical trials." In: *Cancer* 107.7 Suppl, pp. 1645–1655.

Bodmer, W. F. et al. (1987). "Localization of the gene for familial adenomatous polyposis on chromosome 5." In: *Nature* 328.6131, pp. 614–616.

Boland, C. R. et al. (1995). "Microallelotyping Defines the Sequence and Tempo of Allelic Losses at Tumor-Suppressor Gene Loci During Colorectal-Cancer Progression". In: *Nature medicine* 1.9, pp. 902–909.

Boland, C. R. et al. (1998). "A National Cancer Institute Workshop on Microsatellite Instability for Cancer Detection and Familial Predisposition: Development of International Criteria for the Determination of Microsatellite Instability in Colorectal Cancer". In: *Cancer research* 58.22, pp. 5248–5257.

Boland, P. M., M. B. Yurgelun, and C. R. Boland (2018). "Recent progress in Lynch syndrome and other familial colorectal cancer syndromes". In: *Ca-a Cancer Journal for Clinicians* 68.3, pp. 217–231.

Borodulin, K. et al. (2017). "Cohort Profile: The National FINRISK Study." In: *International journal of epidemiology* 47.3, pp. 696–696i.

Botteri, E. et al. (2008). "Smoking and colorectal cancer: a meta-analysis." In: *JAMA* 300.23, pp. 2765–2778.

Bouvard, V. et al. (2015). “Carcinogenicity of consumption of red and processed meat.” In: *The Lancet. Oncology* 16.16, pp. 1599–1600.

Brensinger, J. D. et al. (1998). “Variable phenotype of familial adenomatous polyposis in pedigrees with 3’ mutation in the APC gene.” In: *Gut* 43.4, pp. 548–552.

Broderick, P. et al. (2007). “A genome-wide association study shows that common alleles of SMAD7 influence colorectal cancer risk.” In: *Nature genetics* 39.11, pp. 1315–1317.

Broeke, S. W. ten et al. (2015). “Lynch syndrome caused by germline PMS2 mutations: delineating the cancer risk.” In: *Journal of clinical oncology : official journal of the American Society of Clinical Oncology* 33.4, pp. 319–325.

Bronner, C. E. et al. (1994). “Mutation in the DNA mismatch repair gene homologue hMLH1 is associated with hereditary non-polyposis colon cancer.” In: *Nature* 368.6468, pp. 258–261.

Brosens, L. A. A. et al. (2007). “Risk of colorectal cancer in juvenile polyposis”. In: *Gut* 56.7, pp. 965–967.

Bülow, S (2003). “Results of national registration of familial adenomatous polyposis.” In: *Gut* 52.5, pp. 742–746.

Burn, J. et al. (2011). “Long-term effect of aspirin on cancer risk in carriers of hereditary colorectal cancer: an analysis from the CAPP2 randomised controlled trial”. In: *Lancet (London, England)* 378.9809, pp. 2081–2087.

Bussey, H. J. R. (1975). *Familial polyposis coli: family studies, histopathology, differential diagnosis, and results of treatment.* Johns Hopkins University Press.

Calva, D. and J. R. Howe (2008). “Hamartomatous polyposis syndromes.” In: *The Surgical clinics of North America* 88.4, pp. 779–817–vii.

Chapelle, A. de la (2005). “The incidence of Lynch syndrome.” In: *Familial cancer* 4.3, pp. 233–237.

Chapuis, P. H. et al. (1985). “A multivariate analysis of clinical and pathological variables in prognosis after resection of large bowel cancer.” In: *The British journal of surgery* 72.9, pp. 698–702.

Chatterjee, N. et al. (2013). “Projecting the performance of risk prediction based on polygenic analyses of genome-wide association studies”. In: *Nature genetics* 45.4, pp. 400–405.

Chen, H. S. and S. M. Sheen-Chen (2000). “Obstruction and perforation in colorectal adenocarcinoma: an analysis of prognosis and current trends.” In: *Surgery* 127.4, pp. 370–376.

Chenevix-Trench, G. et al. (2002). “Dominant negative ATM mutations in breast cancer families.” In: *JNCI: Journal of the National Cancer Institute* 94.3, pp. 205–215.

Cheng, T. H. T. et al. (2015). “Meta-analysis of genome-wide association studies identifies common susceptibility polymorphisms for colorectal and endometrial cancer near SH2B3 and TSHZ1.” In: *Scientific reports* 5, p. 17369.

Cho, E. et al. (2004). “Alcohol Intake and Colorectal Cancer: A Pooled Analysis of 8 Cohort Studies”. In: *Annals of internal medicine* 140.8, pp. 603–613.

COGENT Study (2008). “Meta-analysis of genome-wide association data identifies four new susceptibility loci for colorectal cancer.” In: *Nature genetics* 40.12, pp. 1426–1435.

Conti, E and E Izaurralde (2005). “Nonsense-mediated mRNA decay: molecular insights and mechanistic variations across species”. In: *Current Opinion in Cell Biology* 17.3, pp. 316–325.

Croce, C. M. (2008). “Oncogenes and cancer.” In: *The New England journal of medicine* 358.5, pp. 502–511.

Dawn Teare, M and J. H. Barrett (2005). “Genetic linkage studies.” In: *Lancet (London, England)* 366.9490, pp. 1036–1044.

Deng, L. et al. (2012). “Diabetes mellitus and the incidence of colorectal cancer: an updated systematic review and meta-analysis.” In: *Digestive diseases and sciences* 57.6, pp. 1576–1585.

Domingo, E et al. (2004). “BRAF screening as a low-cost effective strategy for simplifying HNPCC genetic testing.” In: *Journal of medical genetics* 41.9, pp. 664–668.

Domingo, E. et al. (2016). “Somatic POLE proofreading domain mutation, immune response, and prognosis in colorectal cancer: a retrospective, pooled biomarker study.” In: *The Lancet. Gastroenterology & hepatology* 1.3, pp. 207–216.

Dunlop, M. G. et al. (2012). “Common variation near CDKN1A, POLD3 and SHROOM2 influences colorectal cancer risk.” In: *Nature genetics* 44.7, pp. 770–776.

Elsayed, F. A. et al. (2015). “Germline variants in POLE are associated with early onset mismatch repair deficient colorectal cancer.” In: *European journal of human genetics* 23.8, pp. 1080–1084.

Esteller, M. (2011). “Non-coding RNAs in human disease.” In: *Nature reviews. Genetics* 12.12, pp. 861–874.

Evangelou, E. and J. P. A. Ioannidis (2013). “Meta-analysis methods for genome-wide association studies and beyond.” In: *Nature reviews. Genetics* 14.6, pp. 379–389.

Fearon, E. R. and B. Vogelstein (1990). “A genetic model for colorectal tumorigenesis”. In: *Cell* 61.5, pp. 759–767.

Federico, A. et al. (2017). “Pan-Cancer Mutational and Transcriptional Analysis of the Integrator Complex.” In: *International journal of molecular sciences* 18.5, p. 936.

Ferlay, J. et al. (2015). “Cancer incidence and mortality worldwide: Sources, methods and major patterns in GLOBOCAN 2012”. In: *International journal of cancer* 136.5, E359–E386.

Fidler, M. M. et al. (2017). “Cancer incidence and mortality among young adults aged 20-39 years worldwide in 2012: a population-based study.” In: *The Lancet. Oncology* 18.12, pp. 1579–1589.

Fishel, R et al. (1993). “The human mutator gene homolog MSH2 and its association with hereditary nonpolyposis colon cancer.” In: *Cell* 75.5, pp. 1027–1038.

Fletcher, O. and R. S. Houlston (2010). “Architecture of inherited susceptibility to common cancer.” In: *Nature reviews. Cancer* 10.5, pp. 353–361.

Forrester, K et al. (1987). “Detection of High-Incidence of K-Ras Oncogenes During Human-Colon Tumorigenesis”. In: *Nature* 327.6120, pp. 298–303.

Frampton, M. J. E. et al. (2016). “Implications of polygenic risk for personalised colorectal cancer screening.” In: *Annals of oncology : official journal of the European Society for Medical Oncology* 27.3, pp. 429–434.

Freedman, M. L. et al. (2011). “Principles for the post-GWAS functional characterization of cancer risk loci.” In: *Nature genetics* 43.6, pp. 513–518.

Friend, S. H. et al. (1986). “A human DNA segment with properties of the gene that predisposes to retinoblastoma and osteosarcoma.” In: *Nature* 323.6089, pp. 643–646.

Galiatsatos, P. and W. D. Foulkes (2006). “Familial adenomatous polyposis.” In: *American journal of gastroenterology* 101.2, pp. 385–398.

Gardner, E. J. (1951). “A Genetic and Clinical Study of Intestinal Polyposis, a Predisposing Factor for Carcinoma of the Colon and Rectum”. In: *Am. J. Hum. Genet.* 3.2, pp. 167–176.

Gharahkhani, P. et al. (2016). “Genome-wide association studies in oesophageal adenocarcinoma and Barrett’s oesophagus: a large-scale meta-analysis.” In: *The Lancet. Oncology* 17.10, pp. 1363–1373.

Giacinti, C and A Giordano (2006). “RB and cell cycle progression”. In: *Oncogene* 25.38, pp. 5220–5227.

Giardiello, F. M. et al. (2000). “Very high risk of cancer in familial Peutz-Jeghers syndrome”. In: *Gastroenterology* 119.6, pp. 1447–1453.

Giardiello, F. M. et al. (2014). “Guidelines on genetic evaluation and management of Lynch syndrome: a consensus statement by the US Multi-Society Task Force on Colorectal Cancer.” In: *Diseases of the colon and rectum* 57.8, pp. 1025–1048.

Gibson, G. (2012). “Rare and common variants: twenty arguments.” In: *Nature reviews. Genetics* 13.2, pp. 135–145.

Goelz, S. E. et al. (1985). “Hypomethylation of DNA from benign and malignant human colon neoplasms.” In: *Science* 228.4696, pp. 187–190.

Goldfeder, R. L. et al. (2017). “Human Genome Sequencing at the Population Scale: A Primer on High-Throughput DNA Sequencing and Analysis.” In: *American journal of epidemiology* 186.8, pp. 1000–1009.

Goldgar, D. E. et al. (1994). “Systematic Population-Based Assessment of Cancer Risk in First-Degree Relatives of Cancer Probands”. In: *JNCI: Journal of the National Cancer Institute* 86.21, pp. 1600–1608.

Goldin, L. R. et al. (2004). “Familial risk of lymphoproliferative tumors in families of patients with chronic lymphocytic leukemia: results from the Swedish Family-Cancer Database.” In: *Blood* 104.6, pp. 1850–1854.

Graff, R. E. et al. (2017). “Familial Risk and Heritability of Colorectal Cancer in the Nordic Twin Study of Cancer.” In: *Clinical gastroenterology and hepatology : the official clinical practice journal of the American Gastroenterological Association* 15.8, pp. 1256–1264.

Grainger, D. J. et al. (1999). “Genetic control of the circulating concentration of transforming growth factor type beta1.” In: *Human Molecular Genetics* 8.1, pp. 93–97.

Griffin, P. M. et al. (1991). “Adenocarcinomas of the colon and rectum in persons under 40 years old. A population-based study.” In: *Gastroenterology* 100.4, pp. 1033–1040.

Groden, J et al. (1991). “Identification and characterization of the familial adenomatous polyposis coli gene.” In: *Cell* 66.3, pp. 589–600.

Guilford, P et al. (1998). “E-cadherin germline mutations in familial gastric cancer.” In: *Nature* 392.6674, pp. 402–405.

Guinney, J. et al. (2015). “The consensus molecular subtypes of colorectal cancer”. In: *Nature medicine* 21.11, pp. 1350–1356.

Gylfe, A. E. et al. (2013). “Eleven candidate susceptibility genes for common familial colorectal cancer.” In: *PLoS genetics* 9.10, e1003876.

Haldane, J. B. S. and C. A. B. Smith (1947). “A new estimate of the linkage between the genes for colourblindness and haemophilia in man.” In: *Annals of eugenics* 14.pt 1, pp. 10–31.

Hampel, H. et al. (2005). “Screening for the Lynch syndrome (hereditary nonpolyposis colorectal cancer).” In: *The New England journal of medicine* 352.18, pp. 1851–1860.

Hanahan, D. and R. A. Weinberg (2011). “Hallmarks of cancer: the next generation.” In: *Cell* 144.5, pp. 646–674.

Harris, R. S. (2013). “Cancer mutation signatures, DNA damage mechanisms, and potential clinical implications”. In: *Genome Medicine* 5.9, p. 87.

Hatzikotoulas, K., A. Gilly, and E. Zeggini (2014). “Using population isolates in genetic association studies.” In: *Briefings in Functional Genomics* 13.5, pp. 371–377.

Hause, R. J. et al. (2016). “Classification and characterization of microsatellite instability across 18 cancer types.” In: *Nature medicine* 22.11, pp. 1342–1350.

Heald, B. et al. (2010). “Frequent gastrointestinal polyps and colorectal adenocarcinomas in a prospective series of PTEN mutation carriers.” In: *Gastroenterology* 139.6, pp. 1927–1933.

Heald, B. et al. (2013). “Implementation of universal microsatellite instability and immunohistochemistry screening for diagnosing lynch syndrome in a large academic medical center.” In: *Journal of clinical oncology : official journal of the American Society of Clinical Oncology* 31.10, pp. 1336–1340.

Hearle, N. et al. (2006). “Frequency and Spectrum of Cancers in the Peutz-Jeghers Syndrome”. In: *Clinical cancer research: an official journal of the american association for cancer research* 12.10, pp. 3209–3215.

Helgason, H. et al. (2015). “Loss-of-function variants in ATM confer risk of gastric cancer.” In: *Nature genetics* 47.8, pp. 906–910.

Helwig, E. B. (1946). “Adenomas of the large intestine in children.” In: *Am. J. Dis. Child.* 72, pp. 289–295.

Hemminki, A et al. (1997). “Localization of a susceptibility locus for Peutz-Jeghers syndrome to 19p using comparative genomic hybridization and targeted linkage analysis.” In: *Nature genetics* 15.1, pp. 87–90.

Hemminki, A et al. (1998a). “A serine/threonine kinase gene defective in Peutz-Jeghers syndrome.” In: *Nature* 391.6663, pp. 184–187.

Hemminki, K, P Vaittinen, and P Kyyrönen (1998b). “Age-specific familial risks in common cancers of the offspring.” In: *International journal of cancer* 78.2, pp. 172–175.

Heppner Goss, K et al. (2002). “Attenuated APC alleles produce functional protein from internal translation initiation.” In: *Proceedings of the National Academy of Sciences of the United States of America* 99.12, pp. 8161–8166.

Herman, J. G. et al. (1994). “Silencing of the VHL tumor-suppressor gene by DNA methylation in renal carcinoma.” In: *Proceedings of the National Academy of Sciences of the United States of America* 91.21, pp. 9700–9704.

Herrera, L et al. (1986). “Gardner syndrome in a man with an interstitial deletion of 5q.” In: 25.3, pp. 473–476.

Hobert, J. A. and C. Eng (2009). “PTEN hamartoma tumor syndrome: an overview.” In: *Genetics in medicine : official journal of the American College of Medical Genetics* 11.10, pp. 687–694.

Hofstra, R. M. et al. (1994). “A mutation in the RET proto-oncogene associated with multiple endocrine neoplasia type 2B and sporadic medullary thyroid carcinoma.” In: *Nature* 367.6461, pp. 375–376.

Hogan, J. et al. (2015). “Lymphovascular invasion: a comprehensive appraisal in colon and rectal adenocarcinoma.” In: *Diseases of the colon and rectum* 58.6, pp. 547–555.

Holst, S von et al. (2010). “Association studies on 11 published colorectal cancer risk loci”. In: *British journal of cancer* 103.4, pp. 575–580.

Houlston, R. S. and I. P. Tomlinson (2001). “Polymorphisms and colorectal tumor risk.” In: *Gastroenterology* 121.2, pp. 282–301.

Houlston, R. S. et al. (2010). “Meta-analysis of three genome-wide association studies identifies susceptibility loci for colorectal cancer at 1q41, 3q26.2, 12q13.13 and 20q13.33.” In: *Nature genetics* 42.11, pp. 973–977.

Howe, J. R. et al. (1998a). “A gene for familial juvenile polyposis maps to chromosome 18q21.1.” In: *Am. J. Hum. Genet.* 62.5, pp. 1129–1136.

Howe, J. R. et al. (1998b). “Mutations in the SMAD4/DPC4 gene in juvenile polyposis.” In: *Science* 280.5366, pp. 1086–1088.

Howe, J. R. et al. (2001). “Germline mutations of the gene encoding bone morphogenetic protein receptor 1A in juvenile polyposis.” In: *Nature genetics* 28.2, pp. 184–187.

Howe, J. R. et al. (2004). “The prevalence of MADH4 and BMPR1A mutations in juvenile polyposis and absence of BMPR2, BMPR1B, and ACVR1 mutations.” In: *Journal of medical genetics* 41.7, pp. 484–491.

Hubbard, J. et al. (2012). “Benefits and adverse events in younger versus older patients receiving adjuvant chemotherapy for colon cancer: findings from the Adjuvant Colon Cancer Endpoints data set.” In: *Journal of clinical oncology : official journal of the American Society of Clinical Oncology* 30.19, pp. 2334–2339.

International HapMap Consortium (2005). “A haplotype map of the human genome.” In: *Nature* 437.7063, pp. 1299–1320.

Jaeger, E. E. M. et al. (2003). “An ancestral Ashkenazi haplotype at the HMPS/CRAC1 locus on 15q13-q14 is associated with hereditary mixed polyposis syndrome.” In: *Am. J. Hum. Genet.* 72.5, pp. 1261–1267.

Jaeger, E. et al. (2008). “Common genetic variants at the CRAC1 (HMPS) locus on chromosome 15q13.3 influence colorectal cancer risk”. In: *Nature genetics* 40.1, pp. 26–28.

Jaeger, E. et al. (2012). “Hereditary mixed polyposis syndrome is caused by a 40-kb upstream duplication that leads to increased and ectopic expression of the BMP antagonist GREM1.” In: *Nature genetics* 44.6, pp. 699–703.

Jamshidi, F., T. O. Nielsen, and D. G. Huntsman (2015). “Cancer genomics: why rare is valuable.” In: *Journal of molecular medicine (Berlin, Germany)* 93.4, pp. 369–381.

Järvinen, H. J. et al. (2000). “Controlled 15-year trial on screening for colorectal cancer in families with hereditary nonpolyposis colorectal cancer.” In: *Gastroenterology* 118.5, pp. 829–834.

Jass, J. R. et al. (1988). “Juvenile polyposis–a precancerous condition.” In: *Histopathology* 13.6, pp. 619–630.

Jeggo, P. A., L. H. Pearl, and A. M. Carr (2016). “DNA repair, genome stability and cancer: a historical perspective.” In: *Nature reviews. Cancer* 16.1, pp. 35–42.

Jeghers, H, V. A. McKusick, and K. H. Katz (1949). “Generalized intestinal polyposis and melanin spots of the oral mucosa, lips and digits; a syndrome of diagnostic significance.” In: *The New England journal of medicine* 241.26, pp. 1031–1036.

Jess, T., C. Rungoe, and L. Peyrin-Biroulet (2012). “Risk of Colorectal Cancer in Patients With Ulcerative Colitis: A Meta-analysis of Population-Based Cohort Studies”. In: *Clinical gastroenterology and hepatology : the official clinical practice journal of the American Gastroenterological Association* 10.6, pp. 639–645.

Jia, W.-H. et al. (2013). “Genome-wide association analyses in East Asians identify new susceptibility loci for colorectal cancer.” In: *Nature genetics* 45.2, pp. 191–196.

Jiao, S. et al. (2014). “Estimating the heritability of colorectal cancer”. In: *Human Molecular Genetics* 23.14, pp. 3898–3905.

Johns, L. E. and R. S. Houlston (2001). “A systematic review and meta-analysis of familial colorectal cancer risk”. In: *American journal of gastroenterology* 96.10, pp. 2992–3003.

Jones, J. C. et al. (2017). “Non-V600 BRAF Mutations Define a Clinically Distinct Molecular Subtype of Metastatic Colorectal Cancer.” In: *Journal of clinical oncology : official journal of the American Society of Clinical Oncology* 35.23, pp. 2624–2630.

Joslyn, G et al. (1991). “Identification of deletion mutations and three new genes at the familial polyposis locus.” In: *Cell* 66.3, pp. 601–613.

Jostins, L. et al. (2012). “Host-microbe interactions have shaped the genetic architecture of inflammatory bowel disease.” In: *Nature* 491.7422, pp. 119–124.

Kakar, S. et al. (2003). “Frequency of loss of hMLH1 expression in colorectal carcinoma increases with advancing age.” In: *Cancer* 97.6, pp. 1421–1427.

Kalady, M. F. et al. (2015). “Defining the adenoma burden in lynch syndrome.” In: *Diseases of the colon and rectum* 58.4, pp. 388–392.

Kalia, S. S. et al. (2017). “Recommendations for reporting of secondary findings in clinical exome and genome sequencing, 2016 update (ACMG SF v2.0): a policy statement of the American College of Medical Genetics and Genomics.” In: *Genetics in medicine : official journal of the American College of Medical Genetics* 19.2, pp. 249–255.

Kandoth, C. et al. (2013). “Mutational landscape and significance across 12 major cancer types.” In: *Nature* 502.7471, pp. 333–339.

Kane, M. F. et al. (1997). “Methylation of the hMLH1 promoter correlates with lack of expression of hMLH1 in sporadic colon tumors and mismatch repair-defective human tumor cell lines.” In: *Cancer research* 57.5, pp. 808–811.

Katainen, R. et al. (2015). “CTCF/cohesin-binding sites are frequently mutated in cancer.” In: *Nature genetics* 47.7, pp. 818–821.

Katainen, R. et al. (2017). “BasePlayer: Versatile Analysis Software For Large-Scale Genomic Variant Discovery”. In:

Kim, H et al. (1994). “Clinical and pathological characteristics of sporadic colorectal carcinomas with DNA replication errors in microsatellite sequences.” In: *The American journal of pathology* 145.1, pp. 148–156.

Kim, J. K. and J. A. Diehl (2009). “Nuclear cyclin D1: an oncogenic driver in human cancer.” In: *Journal of cellular physiology* 220.2, pp. 292–296.

Kinzler, K. W. et al. (1991). “Identification of FAP locus genes from chromosome 5q21.” In: *Science* 253.5020, pp. 661–665.

Kinzler, K. W. and B Vogelstein (1996). “Lessons from hereditary colorectal cancer.” In: *Cell* 87.2, pp. 159–170.

— (1997). “Cancer-susceptibility genes. Gatekeepers and caretakers.” In: *Nature* 386.6627, pp. 761–763.

Klaus, C. et al. (2014). “Modulating effects of acyl-CoA synthetase 5-derived mitochondrial Wnt2B palmitoylation on intestinal Wnt activity.” In: *World journal of gastroenterology* 20.40, pp. 14855–14864.

Kloor, M. et al. (2012). “Prevalence of mismatch repair-deficient crypt foci in Lynch syndrome: A pathological study”. In: *The Lancet. Oncology* 13.6, pp. 598–606.

Knudsen, A. L., M.-L. Bisgaard, and S. Bülow (2003). “Attenuated familial adenomatous polyposis (AFAP). A review of the literature.” In: *Familial cancer* 2.1, pp. 43–55.

Knudson, A. G. (1971). “Mutation and cancer: statistical study of retinoblastoma.” In: *Proceedings of the National Academy of Sciences of the United States of America* 68.4, pp. 820–823.

Kondelin, J. et al. (2017). “Comprehensive Evaluation of Protein Coding Mononucleotide Microsatellites in Microsatellite-Unstable Colorectal Cancer.” In: *Cancer research* 77.15, pp. 4078–4088.

Koushik, A. et al. (2007). “Fruits, vegetables, and colon cancer risk in a pooled analysis of 14 cohort studies.” In: *Journal of the National Cancer Institute* 99.19, pp. 1471–1483.

Kovacs, M. E. et al. (2009). “Deletions Removing the Last Exon of TACSTD1 Constitute a Distinct Class of Mutations Predisposing to Lynch Syndrome”. In: *Human Mutation* 30.2, pp. 197–203.

Kraft, P. (2017). “Fine Tuning the Risk of Hereditary Cancer Using Genome-Wide Association Studies”. In: *Journal of Clinical Oncology* 35.20, pp. 2224–2225.

Kraft, P., E. Zeggini, and J. P. A. Ioannidis (2009). “Replication in genome-wide association studies.” In: *Statistical science : a review journal of the Institute of Mathematical Statistics* 24.4, pp. 561–573.

Kristiansson, K., J. Naukkarinen, and L. Peltonen (2008). “Isolated populations and complex disease gene identification.” In: *Genome biology* 9.8, p. 109.

Kryukov, G. V., L. A. Pennacchio, and S. R. Sunyaev (2007). “Most rare missense alleles are deleterious in humans: implications for complex disease and association studies.” In: *Am. J. Hum. Genet.* 80.4, pp. 727–739.

Kumar, V., A. K. Abbas, and J. C. Aster (2014). *Robbins and Cotran Pathologic Basis of Disease*. Saunders.

Kunkel, T. A. and D. A. Erie (2005). “DNA mismatch repair.” In: *Annual review of biochemistry* 74.1, pp. 681–710.

Laken, S. J. et al. (1997). “Familial colorectal cancer in Ashkenazim due to a hypermutable tract in APC.” In: *Nature genetics* 17.1, pp. 79–83.

Lalani, S. R. et al. (2013). “MCTP2 is a dosage-sensitive gene required for cardiac outflow tract development.” In: *Human Molecular Genetics* 22.21, pp. 4339–4348.

Lammi, L. et al. (2004). “Mutations in AXIN2 cause familial tooth agenesis and predispose to colorectal cancer.” In: *Am. J. Hum. Genet.* 74.5, pp. 1043–1050.

Larsson, S. C. and A. Wolk (2007). “Obesity and colon and rectal cancer risk: A meta-analysis of prospective studies”. In: *American journal of clinical nutrition* 86.3, pp. 556–565.

Lawrence, M. S. et al. (2014). “Discovery and saturation analysis of cancer genes across 21 tumour types.” In: *Nature* 505.7484, pp. 495–501.

Le, D. T. et al. (2017). “Programmed death-1 blockade in mismatch repair deficient colorectal cancer.” In: *Journal of Clinical Oncology*.

Leach, F. S. et al. (1993). “Mutations of a mutS homolog in hereditary nonpolyposis colorectal cancer.” In: *Cell* 75.6, pp. 1215–1225.

Lee, I.-M. et al. (2012). “Effect of physical inactivity on major non-communicable diseases worldwide: An analysis of burden of disease and life expectancy”. In: *The Lancet* 380.9838, pp. 219–229.

Leggett, B. and V. Whitehall (2010). “Role of the Serrated Pathway in Colorectal Cancer Pathogenesis”. In: *Gastroenterology* 138.6, pp. 2088–2100.

Lek, M. et al. (2016). “Analysis of protein-coding genetic variation in 60,706 humans.” In: *Nature* 536.7616, pp. 285–291.

Liang, J. et al. (2013). “APC polymorphisms and the risk of colorectal neoplasia: a HuGE review and meta-analysis.” In: *American journal of epidemiology* 177.11, pp. 1169–1179.

Lichtenstein, P et al. (2000). "Environmental and heritable factors in the causation of cancer–analyses of cohorts of twins from Sweden, Denmark, and Finland." In: *The New England journal of medicine* 343.2, pp. 78–85.

Liebig, C. et al. (2009). "Perineural invasion is an independent predictor of outcome in colorectal cancer." In: *Journal of clinical oncology : official journal of the American Society of Clinical Oncology* 27.31, pp. 5131–5137.

Lier, M. G. F. van et al. (2010). "High Cancer Risk in Peutz-Jeghers Syndrome: A Systematic Review and Surveillance Recommendations". In: *American journal of gastroenterology* 105.6, pp. 1258–1264.

Ligtenberg, M. J. L. et al. (2009). "Heritable somatic methylation and inactivation of MSH2 in families with Lynch syndrome due to deletion of the 3' exons of TACSTD1." In: *Nature genetics* 41.1, pp. 112–117.

Lin, E. I. et al. (2015). "Mutational profiling of colorectal cancers with microsatellite instability." In: *Oncotarget* 6.39, pp. 42334–42344.

Lin, J. S. et al. (2016). "Screening for Colorectal Cancer: Updated Evidence Report and Systematic Review for the US Preventive Services Task Force." In: *JAMA* 315.23, pp. 2576–2594.

Liu, C., Q.-S. Wang, and Y.-J. Wang (2012). "The CHEK2 I157T variant and colorectal cancer susceptibility: a systematic review and meta-analysis." In: *Asian Pacific journal of cancer prevention : APJCP* 13.5, pp. 2051–2055.

Liu, X et al. (2000). "Molecular cloning and chromosomal mapping of a candidate cytokine gene selectively expressed in human CD34+ cells." In: *Genomics* 65.3, pp. 283–292.

Lubbe, S. J. et al. (2009). "Clinical implications of the colorectal cancer risk associated with MUTYH mutation." In: *Journal of clinical oncology : official journal of the American Society of Clinical Oncology* 27.24, pp. 3975–3980.

Luebeck, E. G. and S. H. Moolgavkar (2002). "Multistage carcinogenesis and the incidence of colorectal cancer". In: *Proceedings of the National Academy of Sciences of the United States of America* 99.23, pp. 15095–15100.

Lutgens, M. W. M. D. et al. (2013). "Declining risk of colorectal cancer in inflammatory bowel disease: an updated meta-analysis of population-based cohort studies." In: *Inflammatory bowel diseases* 19.4, pp. 789–799.

Lyons, J. P. et al. (2004). "Wnt-4 activates the canonical beta-catenin-mediated Wnt pathway and binds Frizzled-6 CRD: functional implications of Wnt/beta-catenin activity in kidney epithelial cells." In: *Experimental cell research* 298.2, pp. 369–387.

MacArthur, D. G. et al. (2012). "A systematic survey of loss-of-function variants in human protein-coding genes." In: *Science* 335.6070, pp. 823–828.

Malesci, A. et al. (2007). "Reduced likelihood of metastases in patients with microsatellite-unstable colorectal cancer." In: *Clinical cancer research: an of-*

ficial journal of the american association for cancer research 13.13, pp. 3831–3839.

Malkin, D. (2011). “Li-Fraumeni syndrome.” In: *Genes & cancer* 2.4, pp. 475–484.

Mandelker, D. et al. (2017). “Mutation Detection in Patients With Advanced Cancer by Universal Sequencing of Cancer-Related Genes in Tumor and Normal DNA vs Guideline-Based Germline Testing.” In: *JAMA* 318.9, pp. 825–835.

Martincorena, I. et al. (2015). “Tumor evolution. High burden and pervasive positive selection of somatic mutations in normal human skin.” In: *Science* 348.6237, pp. 880–886.

Maughan, T. S. et al. (2011). “Addition of cetuximab to oxaliplatin-based first-line combination chemotherapy for treatment of advanced colorectal cancer: results of the randomised phase 3 MRC COIN trial.” In: *Lancet (London, England)* 377.9783, pp. 2103–2114.

McCarthy, S. et al. (2016). “A reference panel of 64,976 haplotypes for genotype imputation.” In: *Nature genetics* 48.10, pp. 1279–1283.

Meer, J. W. van der et al. (1993). “Colorectal cancer in patients with X-linked agammaglobulinaemia.” In: *Lancet (London, England)* 341.8858, pp. 1439–1440.

Mehlen, P and A Puisieux (2006). “Metastasis: a question of life or death”. In: *Nature reviews. Cancer* 6.6, pp. 449–458.

Merlo, A et al. (1995). “5’ CpG island methylation is associated with transcriptional silencing of the tumour suppressor p16/CDKN2/MTS1 in human cancers.” In: *Nature medicine* 1.7, pp. 686–692.

Metzker, M. L. (2010). “Sequencing technologies - the next generation.” In: *Nature reviews. Genetics* 11.1, pp. 31–46.

Miller, K. D. et al. (2016). “Cancer treatment and survivorship statistics, 2016”. In: *Ca-a Cancer Journal for Clinicians* 66.4, pp. 271–289.

Miller, N. A. et al. (2015). “A 26-hour system of highly sensitive whole genome sequencing for emergency management of genetic diseases.” In: *Genome Medicine* 7.1, p. 100.

Miyaki, M et al. (1997). “Germline mutation of MSH6 as the cause of hereditary nonpolyposis colorectal cancer.” In: *Nature genetics* 17.3, pp. 271–272.

Modest, D. P. et al. (2016). “Outcome according to KRAS-, NRAS- and BRAF-mutation as well as KRAS mutation variants: pooled analysis of five randomized trials in metastatic colorectal cancer by the AIO colorectal cancer study group.” In: *Annals of oncology : official journal of the European Society for Medical Oncology* 27.9, pp. 1746–1753.

Møller, P. et al. (2017). “Cancer risk and survival in pathMMR carriers by gene and gender up to 75 years of age: a report from the Prospective Lynch Syndrome Database.” In: *Gut*.

Moreira, L. et al. (2012). “Identification of Lynch syndrome among patients with colorectal cancer.” In: *JAMA* 308.15, pp. 1555–1565.

Morin, P. J. et al. (1997). “Activation of beta-catenin-Tcf signaling in colon cancer by mutations in beta-catenin or APC.” In: *Science* 275.5307, pp. 1787–1790.

Mork, M. E. et al. (2015). “High Prevalence of Hereditary Cancer Syndromes in Adolescents and Young Adults With Colorectal Cancer.” In: *Journal of clinical oncology : official journal of the American Society of Clinical Oncology* 33.31, pp. 3544–3549.

Mucci, L. A. et al. (2016). “Familial Risk and Heritability of Cancer Among Twins in Nordic Countries.” In: *JAMA* 315.1, pp. 68–76.

Mulligan, L. M. et al. (1993). “Germ-line mutations of the RET proto-oncogene in multiple endocrine neoplasia type 2A.” In: *Nature* 363.6428, pp. 458–460.

Murday, V and J Slack (1989). “Inherited disorders associated with colorectal cancer.” In: *Cancer surveys* 8.1, pp. 139–157.

Murff, H. J., D. R. Spigel, and S. Syngal (2004). “Does this patient have a family history of cancer? An evidence-based analysis of the accuracy of family cancer history.” In: *JAMA* 292.12, pp. 1480–1489.

Muzny, D. M. et al. (2012). “Comprehensive molecular characterization of human colon and rectal cancer”. In: *Nature* 487.7407, pp. 330–337.

Nagy, R., K. Sweet, and C. Eng (2004). “Highly penetrant hereditary cancer syndromes.” In: *Oncogene* 23.38, pp. 6445–6470.

Näslund-Koch, C., B. G. Nordestgaard, and S. E. Bojesen (2016). “Increased Risk for Other Cancers in Addition to Breast Cancer for CHEK2*1100delC Heterozygotes Estimated From the Copenhagen General Population Study.” In: *Journal of clinical oncology : official journal of the American Society of Clinical Oncology* 34.11, pp. 1208–1216.

Negrini, S., V. G. Gorgoulis, and T. D. Halazonetis (2010). “Genomic instability–an evolving hallmark of cancer.” In: *Nature reviews. Molecular cell biology* 11.3, pp. 220–228.

Nelson, M. R. et al. (2015). “The support of human genetic evidence for approved drug indications.” In: *Nature genetics* 47.8, pp. 856–860.

Nicolaides, N. C. et al. (1994). “Mutations of two PMS homologues in hereditary nonpolyposis colon cancer.” In: *Nature* 371.6492, pp. 75–80.

Nielsen, M et al. (2005). “Multiplicity in polyp count and extracolonic manifestations in 40 Dutch patients with MYH associated polyposis coli (MAP).” In: *Journal of medical genetics* 42.9, e54–e54.

Nishisho, I et al. (1991). “Mutations of chromosome 5q21 genes in FAP and colorectal cancer patients.” In: *Science* 253.5020, pp. 665–669.

Nissan, A et al. (1999). “Signet-ring cell carcinoma of the colon and rectum: a matched control study.” In: *Diseases of the colon and rectum* 42.9, pp. 1176–1180.

Nordling, C. O. (1953). “A new theory on the cancer-inducing mechanism”. In: *British journal of cancer* 7.1, pp. 68–72.

Nowell, P. C. (1976). “The clonal evolution of tumor cell populations.” In: *Science* 194.4260, pp. 23–28.

Nugent, K. P. et al. (1993). “Solitary juvenile polyps: not a marker for subsequent malignancy.” In: *Gastroenterology* 105.3, pp. 698–700.

Nyström-Lahti, M et al. (1995). “Founding mutations and Alu-mediated recombination in hereditary colon cancer.” In: *Nature medicine* 1.11, pp. 1203–1206.

O’Connell, J. B. et al. (2004a). “Colorectal cancer in the young”. In: *American Journal of Surgery* 187.3, pp. 343–348.

O’Connell, J. B. et al. (2004b). “Do young colon cancer patients have worse outcomes?” In: *World Journal of Surgery* 28.6, pp. 558–562.

Ohue, M et al. (1994). “A Frequent Alteration of P53 Gene in Carcinoma in Adenoma of Colon”. In: *Cancer research* 54.17, pp. 4798–4804.

O’Malley, M. et al. (2012). “The prevalence of hereditary hemorrhagic telangiectasia in juvenile polyposis syndrome.” In: *Diseases of the colon and rectum* 55.8, pp. 886–892.

Orlando, G. et al. (2016). “Variation at 2q35 (PNKD and TMBIM1) influences colorectal cancer risk and identifies a pleiotropic effect with inflammatory bowel disease.” In: *Human Molecular Genetics* 25.11, pp. 2349–2359.

Palles, C. et al. (2013). “Germline mutations affecting the proofreading domains of POLE and POLD1 predispose to colorectal adenomas and carcinomas.” In: *Nature genetics* 45.2, pp. 136–144.

Papadopoulos, N et al. (1994). “Mutation of a mutL homolog in hereditary colon cancer.” In: *Science* 263.5153, pp. 1625–1629.

Parkin, D. M., L Boyd, and L. C. Walker (2011). “16. The fraction of cancer attributable to lifestyle and environmental factors in the UK in 2010.” In: *British journal of cancer* 105 Suppl 2, S77–81.

Parsons, D. W. et al. (2014). “Clinical tumor sequencing: an incidental casualty of the American College of Medical Genetics and Genomics recommendations for reporting of incidental findings.” In: *Journal of clinical oncology : official journal of the American Society of Clinical Oncology* 32.21, pp. 2203–2205.

Pasaniuc, B. and A. L. Price (2017). “Dissecting the genetics of complex traits using summary association statistics.” In: *Nature reviews. Genetics* 18.2, pp. 117–127.

Pearlman, R. et al. (2017). “Prevalence and Spectrum of Germline Cancer Susceptibility Gene Mutations Among Patients With Early-Onset Colorectal Cancer.” In: *JAMA oncology* 3.4, pp. 464–471.

Pe'er, I. et al. (2008). "Estimation of the multiple testing burden for genomewide association studies of nearly all common variants." In: *Genetic epidemiology* 32.4, pp. 381–385.

Peltomäki, P et al. (1993). "Genetic mapping of a locus predisposing to human colorectal cancer." In: *Science* 260.5109, pp. 810–812.

Peters, U. et al. (2013). "Identification of Genetic Susceptibility Loci for Colorectal Tumors in a Genome-Wide Meta-analysis." In: *Gastroenterology* 144.4, 799–807.e24.

Peutz, J. (1921). "Very remarkable case of familial polyposis of mucous membrane of intestinal tract and nasopharynx accompanied by peculiar pigmentation of skin and mucous membrane". In: *Ned Maandschr Geneeskd* 10, pp. 134–146.

Pitkäniemi, J et al. (2015). "Effectiveness of screening for colorectal cancer with a faecal occult-blood test, in Finland." In: *BMJ open gastroenterology* 2.1, e000034.

Pomerantz, M. M. et al. (2009). "The 8q24 cancer risk variant rs6983267 shows long-range interaction with MYC in colorectal cancer." In: *Nature genetics* 41.8, pp. 882–884.

Ponti, G. and M. Ponz de Leon (2005). "Muir-Torre syndrome." In: *The Lancet. Oncology* 6.12, pp. 980–987.

Powell, S. M. et al. (1992). "Apc Mutations Occur Early During Colorectal Tumorigenesis". In: *Nature* 359.6392, pp. 235–237.

Pukkala, E. et al. (2018). "Nordic Cancer Registries - an overview of their procedures and data comparability." In: *Acta Oncol.* 57.4, pp. 440–455.

Pylayeva-Gupta, Y., E. Grabocka, and D. Bar-Sagi (2011). "RAS oncogenes: weaving a tumorigenic web." In: *Nature reviews. Cancer* 11.11, pp. 761–774.

Rafnar, T. et al. (2009). "Sequence variants at the TERT-CLPTM1L locus associate with many cancer types." In: *Nature genetics* 41.2, pp. 221–227.

Reich, D. E. and E. S. Lander (2001). "On the allelic spectrum of human disease." In: *Trends in genetics : TIG* 17.9, pp. 502–510.

Reinhardt, H. C. and B. Schumacher (2012). "The p53 network: cellular and systemic DNA damage responses in aging and cancer." In: *Trends in genetics : TIG* 28.3, pp. 128–136.

Renwick, A. et al. (2006). "ATM mutations that cause ataxia-telangiectasia are breast cancer susceptibility alleles." In: *Nature genetics* 38.8, pp. 873–875.

Rex, D. K. et al. (2017). *Colorectal Cancer Screening: Recommendations for Physicians and Patients From the U.S. Multi-Society Task Force on Colorectal Cancer.*

Richards, S. et al. (2015). "Standards and guidelines for the interpretation of sequence variants: a joint consensus recommendation of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology." In: *Genetics in medicine : official journal of the American College of Medical Genetics*. Nature Publishing Group, pp. 405–424.

Riley, R. D., J. P. T. Higgins, and J. J. Deeks (2011). "Interpretation of random effects meta-analyses." In: *BMJ (Clinical research ed.)* 342.feb10 2, pp. d549–d549.

Rimoin, D. L., R. E. Pyeritz, and B. Korf (2013). *Emery and Rimoin's Principles and Practice of Medical Genetics*. Academic Press.

Risch, N and K Merikangas (1996). "The future of genetic studies of complex human diseases." In: *Science* 273.5281, pp. 1516–1517.

Rivera, B et al. (2014). "A novel AXIN2 germline variant associated with attenuated FAP without signs of oligondontia or ectodermal dysplasia." In: *European journal of human genetics* 22.3, pp. 423–426.

Rivera, B., E. Castellsague, and I. Bah (2015). "Biallelic NTHL1 Mutations in a Woman with Multiple Primary Tumors". In: *The New England journal of medicine* 373.20, pp. 1985–1986.

Robson, M. E. et al. (2015). "American Society of Clinical Oncology Policy Statement Update: Genetic and Genomic Testing for Cancer Susceptibility." In: *Journal of clinical oncology : official journal of the American Society of Clinical Oncology* 33.31, pp. 3660–3667.

Rosai, J. (2011). *Rosai and Ackerman's Surgical Pathology*. Elsevier.

Ryland, G. L. et al. (2015). "Loss of heterozygosity: what is it good for?" In: *BMC medical genomics* 8.1, p. 45.

Salovaara, R et al. (2000). "Population-based molecular detection of hereditary nonpolyposis colorectal cancer." In: *Journal of Clinical Oncology* 18.11, pp. 2193–2200.

Sampson, J. N. et al. (2015). "Analysis of Heritability and Shared Heritability Based on Genome-Wide Association Studies for Thirteen Cancer Types." In: *Journal of the National Cancer Institute* 107.12, djv279.

Santarosa, M. and A. Ashworth (2004). "Haploinsufficiency for tumour suppressor genes: when you don't need to go all the way." In: *Biochimica et biophysica acta* 1654.2, pp. 105–122.

Sartore-Bianchi, A. et al. (2016). "Dual-targeted therapy with trastuzumab and lapatinib in treatment-refractory, KRAS codon 12/13 wild-type, HER2-positive metastatic colorectal cancer (HERACLES): a proof-of-concept, multicentre, open-label, phase 2 trial." In: *The Lancet. Oncology* 17.6, pp. 738–746.

Savitsky, K et al. (1995). "A single ataxia telangiectasia gene with a product similar to PI-3 kinase." In: *Science* 268.5218, pp. 1749–1753.

Schmit, S. L. et al. (2014). "A novel colorectal cancer risk locus at 4q32.2 identified from an international genome-wide association study." In: *Carcinogenesis* 35.11, pp. 2512–2519.

Schumacher, F. R. et al. (2015). "Genome-wide association study of colorectal cancer identifies six new susceptibility loci." In: *Nature communications* 6, p. 7138.

Seshagiri, S. et al. (2012). "Recurrent R-spondin fusions in colon cancer". In: *Nature* 488.7413, pp. 660–664.

Sham, P. (1998). *Statistics in Human Genetics.* Oxford University Press.

Sieber, O. M. et al. (2003). "Multiple colorectal adenomas, classic adenomatous polyposis, and germ-line mutations in MYH." In: *The New England journal of medicine* 348.9, pp. 791–799.

Siegel, R. L., A. Jemal, and E. M. Ward (2009). "Increase in incidence of colorectal cancer among young men and women in the United States." In: *Cancer epidemiology, biomarkers & prevention : a publication of the American Association for Cancer Research, cosponsored by the American Society of Preventive Oncology* 18.6, pp. 1695–1698.

Siegel, R. L. et al. (2017). "Colorectal Cancer Incidence Patterns in the United States, 1974-2013." In: *Journal of the National Cancer Institute* 109.8, p. 7.

Simon, M. S. et al. (2012). "Estrogen plus progestin and colorectal cancer incidence and mortality." In: *Journal of clinical oncology : official journal of the American Society of Clinical Oncology* 30.32, pp. 3983–3990.

Sinicrope, F. A. et al. (2013). "Prognostic impact of deficient DNA mismatch repair in patients with stage III colon cancer from a randomized trial of FOLFOX-based adjuvant chemotherapy." In: *Journal of clinical oncology : official journal of the American Society of Clinical Oncology* 31.29, pp. 3664–3672.

Skol, A. D. et al. (2006). "Joint analysis is more efficient than replication-based analysis for two-stage genome-wide association studies". In: *Nature genetics* 38.2, pp. 209–213.

Slowik, V. et al. (2015). "Desmoid tumors complicating Familial Adenomatous Polyposis: a meta-analysis mutation spectrum of affected individuals." In: *BMC gastroenterology* 15.1, p. 84.

Sneddon, J. B. et al. (2006). "Bone morphogenetic protein antagonist gremlin 1 is widely expressed by cancer-associated stromal cells and can promote tumor cell proliferation." In: *Proceedings of the National Academy of Sciences of the United States of America* 103.40, pp. 14842–14847.

Spier, I. et al. (2015). "Frequency and phenotypic spectrum of germline mutations in POLE and seven other polymerase genes in 266 patients with colorectal adenomas and carcinomas." In: *International journal of cancer* 137.2, pp. 320–331.

Sud, A., B. Kinnersley, and R. S. Houlston (2017). "Genome-wide association studies of cancer: current insights and future perspectives." In: *Nature reviews. Cancer* 17.11, pp. 692–704.

Sulem, P. et al. (2015). “Identification of a large set of rare complete human knockouts.” In: *Nature genetics* 47.5, pp. 448–452.

Sun, Y. et al. (2015). “Next-generation diagnostics: gene panel, exome, or whole genome?” In: *Human Mutation* 36.6, pp. 648–655.

Suraweera, N. et al. (2002). “Evaluation of tumor microsatellite instability using five quasimonomorphic mononucleotide repeats and pentaplex PCR.” In: *Gastroenterology* 123.6, pp. 1804–1811.

Sweet, K. et al. (2005). “Molecular classification of patients with unexplained hamartomatous and hyperplastic polyposis.” In: *JAMA* 294.19, pp. 2465–2473.

Syngal, S. et al. (2015). *ACG clinical guideline: Genetic testing and management of hereditary gastrointestinal cancer syndromes.* Nature Publishing Group.

Taggarshe, D. et al. (2013). “Colorectal cancer: are the ”young” being overlooked?” In: *American Journal of Surgery* 205.3, pp. 312–316.

Tan, M.-H. et al. (2012). “Lifetime cancer risks in individuals with germline PTEN mutations.” In: *Clinical cancer research: an official journal of the american association for cancer research* 18.2, pp. 400–407.

Tenesa, A. et al. (2008). “Genome-wide association scan identifies a colorectal cancer susceptibility locus on 11q23 and replicates risk loci at 8q24 and 18q21.” In: *Nature genetics* 40.5, pp. 631–637.

The Lancet Oncology (2017). “Colorectal cancer: a disease of the young?” In: *The Lancet. Oncology* 18.4, p. 413.

Thibodeau, S. N. et al. (1998). “Microsatellite instability in colorectal cancer: different mutator phenotypes and the principal involvement of hMLH1.” In: *Cancer research* 58.8, pp. 1713–1718.

Thomas, G. et al. (2008). “Multiple loci identified in a genome-wide association study of prostate cancer.” In: *Nature genetics* 40.3, pp. 310–315.

Thompson, D. et al. (2005). “Cancer risks and mortality in heterozygous ATM mutation carriers.” In: *Journal of the National Cancer Institute* 97.11, pp. 813–822.

Thompson, S. L., S. F. Bakhoum, and D. A. Compton (2010). “Mechanisms of Chromosomal Instability”. In: *Current Biology* 20.6, R285–R295.

Tomasetti, C. and B. Vogelstein (2015a). “Cancer etiology. Variation in cancer risk among tissues can be explained by the number of stem cell divisions.” In: *Science* 347.6217, pp. 78–81.

Tomasetti, C. et al. (2015b). “Only three driver gene mutations are required for the development of lung and colorectal cancers.” In: *Proceedings of the National Academy of Sciences of the United States of America* 112.1, pp. 118–123.

Tomlinson, I et al. (1999). “Inherited susceptibility to colorectal adenomas and carcinomas: evidence for a new predisposition gene on 15q14-q22.” In: *Gastroenterology* 116.4, pp. 789–795.

Tomlinson, I. P. M. et al. (2010). “COGENT (COlorectal cancer GENeTics): an international consortium to study the role of polymorphic variation on the risk of colorectal cancer.” In: *British journal of cancer* 102.2, pp. 447–454.

Tomlinson, I. et al. (2007). “A genome-wide association scan of tag SNPs identifies a susceptibility variant for colorectal cancer at 8q24.21.” In: *Nature genetics* 39.8, pp. 984–988.

Tomlinson, I. P. M. et al. (2008). “A genome-wide association study identifies colorectal cancer susceptibility loci on chromosomes 10p14 and 8q23.3.” In: *Nature genetics* 40.5, pp. 623–630.

Tomlinson, I. P. M. et al. (2011). “Multiple common susceptibility variants near BMP pathway loci GREM1, BMP4, and BMP2 explain part of the missing heritability of colorectal cancer.” In: *PLoS genetics* 7.6, e1002105.

Toon, C. W. et al. (2013). “BRAFV600E immunohistochemistry facilitates universal screening of colorectal cancers for Lynch syndrome.” In: *The American journal of surgical pathology* 37.10, pp. 1592–1602.

Torkamani, A., N. E. Wineinger, and E. J. Topol (2018). “The personal and clinical utility of polygenic risk scores.” In: *Nature reviews. Genetics* 135, p. 2091.

Tortorella, M et al. (2000). “The thrombospondin motif of aggrecanase-1 (ADAMTS-4) is critical for aggrecan substrate recognition and cleavage.” In: *The Journal of biological chemistry* 275.33, pp. 25791–25797.

Tuupanen, S. et al. (2008). “Allelic imbalance at rs6983267 suggests selection of the risk allele in somatic colorectal tumor evolution”. In: *Cancer research* 68.1, pp. 14–17.

Tuupanen, S. et al. (2009). “The common colorectal cancer predisposition SNP rs6983267 at chromosome 8q24 confers potential to enhanced Wnt signaling.” In: *Nature genetics* 41.8, pp. 885–890.

Umar, A. et al. (2004). “Revised Bethesda Guidelines for hereditary nonpolyposis colorectal cancer (Lynch syndrome) and microsatellite instability.” In: *Journal of the National Cancer Institute.* NIH Public Access, pp. 261–268.

Van Cutsem, E et al. (2016). “ESMO consensus guidelines for the management of patients with metastatic colorectal cancer”. In: *Annals of oncology : official journal of the European Society for Medical Oncology* 27.8, pp. 1386–1422.

Vasen, H. F. et al. (1999). “New clinical criteria for hereditary nonpolyposis colorectal cancer (HNPCC, Lynch syndrome) proposed by the International Collaborative group on HNPCC.” In: *Gastroenterology.* The Netherlands Foundation for the Detection of Hereditary Tumours, Leiden, The Netherlands., pp. 1453–1456.

Vasen, H. F. A. et al. (2014). *Guidelines for surveillance of individuals with constitutional mismatch repair-deficiency proposed by the European Consortium "Care for CMMR-D" (C4CMMR-D).*

Veale, A. M. et al. (1966). "Juvenile polyposis coli." In: *Journal of medical genetics* 3.1, pp. 5–16.

Visscher, P. M., W. G. Hill, and N. R. Wray (2008). "Heritability in the genomics era–concepts and misconceptions." In: *Nature reviews. Genetics* 9.4, pp. 255–266.

Visscher, P. M. et al. (2017). "10 Years of GWAS Discovery: Biology, Function, and Translation." In: *Am. J. Hum. Genet.* 101.1, pp. 5–22.

Vogelstein, B et al. (1989). "Allelotype of Colorectal Carcinomas". In: *Science* 244.4901, pp. 207–211.

Vogelstein, B. and K. W. Kinzler (2004). "Cancer genes and the pathways they control". In: *Nature medicine* 10.8, pp. 789–799.

Vogelstein, B. et al. (2013). "Cancer genome landscapes." In: *Science* 339.6127, pp. 1546–1558.

Vogt, S. et al. (2009). "Expanded extracolonic tumor spectrum in MUTYH-associated polyposis." In: *Gastroenterology* 137.6, 1976–85.e1–10.

Wang, H. et al. (2014). "Trans-ethnic genome-wide association study of colorectal cancer identifies a new susceptibility locus in VTI1A." In: *Nature communications* 5, p. 4613.

Wang, M. et al. (2016). "Common genetic variation in ETV6 is associated with colorectal cancer susceptibility." In: *Nature communications* 7, p. 11478.

Wang, W. Y. S. et al. (2005). "Genome-wide association studies: theoretical and practical concerns." In: *Nature reviews. Genetics* 6.2, pp. 109–118.

Warthin, A. S. (1931). "Heredity of Carcinoma in Man". In: *Annals of internal medicine* 4.7, pp. 681–696.

Wasan, H. et al. (2014). "Intermittent chemotherapy plus either intermittent or continuous cetuximab for first-line treatment of patients with KRAS wild-type advanced colorectal cancer (COIN-B): a randomised phase 2 trial." In: *The Lancet. Oncology* 15.6, pp. 631–639.

Watson, P. and B. Riley (2005). "The tumor spectrum in the Lynch syndrome." In: *Familial cancer* 4.3, pp. 245–248.

Weren, R. D. A. et al. (2015). "A germline homozygous mutation in the base-excision repair gene NTHL1 causes adenomatous polyposis and colorectal cancer." In: *Nature genetics* 47.6, pp. 668–671.

Whiffin, N. et al. (2014). "Identification of susceptibility loci for colorectal cancer in a genome-wide meta-analysis." In: *Human Molecular Genetics* 23.17, pp. 4729–4737.

Willis, R. A. (1952). *The spread of tumors in the human body.* J. & A. Churchill.

Win, A. K. et al. (2014). “Risk of colorectal cancer for carriers of mutations in MUTYH, with and without a family history of cancer.” In: *Gastroenterology* 146.5, 1208–11.e1–5.

Win, A. K. et al. (2017). “Prevalence and Penetrance of Major Genes and Polygenes for Colorectal Cancer.” In: *Cancer epidemiology, biomarkers & prevention : a publication of the American Association for Cancer Research, cosponsored by the American Society of Preventive Oncology* 26.3, pp. 404–412.

Wong, P. et al. (2006). “Prevalence of early onset colorectal cancer in 397 patients with classic Li-Fraumeni syndrome.” In: *Gastroenterology* 130.1, pp. 73–79.

Wood, A. R. et al. (2014). “Defining the role of common variation in the genomic and biological architecture of adult human height.” In: *Nature genetics* 46.11, pp. 1173–1186.

Wood, L. D. et al. (2007). “The genomic landscapes of human breast and colorectal cancers.” In: *Science* 318.5853, pp. 1108–1113.

Xiang, H.-p. et al. (2011). “Meta-analysis of CHEK2 1100delC variant and colorectal cancer susceptibility.” In: *European journal of cancer (Oxford, England : 1990)* 47.17, pp. 2546–2551.

Yurgelun, M. B. et al. (2015). “Germline TP53 Mutations in Patients With Early-Onset Colorectal Cancer in the Colon Cancer Family Registry”. In: *JAMA oncology* 1.2, pp. 214–221.

Zanke, B. W. et al. (2007). “Genome-wide association scan identifies a colorectal cancer susceptibility locus on chromosome 8q24.” In: *Nature genetics* 39.8, pp. 989–994.

Zehir, A. et al. (2017). “Mutational landscape of metastatic cancer revealed from prospective clinical sequencing of 10,000 patients”. In: *Nature medicine* 23.6, pp. 703–713.

Zeng, C. et al. (2016). “Identification of Susceptibility Loci and Genes for Colorectal Cancer Risk.” In: *Gastroenterology* 150.7, pp. 1633–1645.

Zhang, B. et al. (2014a). “Genome-wide association study identifies a new SMAD7 risk variant associated with colorectal cancer risk in East Asians.” In: *International journal of cancer* 135.4, pp. 948–955.

Zhang, B. et al. (2014b). “Large-scale genetic study in East Asians identifies six new loci associated with colorectal cancer risk.” In: *Nature genetics* 46.6, pp. 533–542.

Zhang, Q et al. (2001). “Nesprins: a novel family of spectrin-repeat-containing proteins that localize to the nuclear membrane in multiple tissues.” In: *Journal of cell science* 114.Pt 24, pp. 4485–4498.

Zhou, X. P. et al. (2001). “Germline mutations in BMPR1A/ALK3 cause a subset of cases of juvenile polyposis syndrome and of Cowden and Bannayan-Riley-Ruvalcaba syndromes.” In: *Am. J. Hum. Genet.* 69.4, pp. 704–711.

Zuk, O. et al. (2014). “Searching for missing heritability: designing rare variant association studies.” In: *Proceedings of the National Academy of Sciences of the United States of America* 111.4, E455–64.

I

Scandinavian Journal of Gastroenterology. 2013; 48: 672–678

informa
healthcare

ORIGINAL ARTICLE

Exome sequencing in diagnostic evaluation of colorectal cancer predisposition in young patients

TOMAS TANSKANEN[1], ALEXANDRA E. GYLFE[1], RIKU KATAINEN[1], MINNA TAIPALE[2,5], LAURA RENKONEN-SINISALO[3], JUKKA-PEKKA MECKLIN[4], HEIKKI JÄRVINEN[3], SARI TUUPANEN[1], OUTI KILPIVAARA[1], PIA VAHTERISTO[1] & LAURI A. AALTONEN[1]

[1]*Department of Medical Genetics, Genome-Scale Biology Program, University of Helsinki, Biomedicum, P.O. Box 63, FIN-00014 University of Helsinki, Helsinki, Finland,* [2]*Genome-Scale Biology Program, Institute of Biomedicine, University of Helsinki, Biomedicum, P.O. Box 63, FIN-00014 University of Helsinki, Helsinki, Finland,* [3]*Department of Surgery, Helsinki University Central Hospital, Hospital District of Helsinki and Uusimaa, Haartmaninkatu 4, Helsinki, Finland,* [4]*Department of Surgery, Jyväskylä Central Hospital, University of Eastern Finland, Keskussairaalantie 19, Jyväskylä, Finland, and* [5]*Science for Life Center, Department of Biosciences and Nutrition, Karolinska Institutet, Box 1031, Solna, Sweden*

Abstract

Objective. Early-onset colorectal cancer (CRC), defined here as age of onset less than 40 years, develops frequently in genetically predisposed individuals. Next-generation sequencing is an increasingly available option in the diagnostic workup of suspected hereditary susceptibility, but little is known about the practical feasibility and additional diagnostic yield of the technology in this patient group. ***Materials and methods.*** We analyzed 38 young CRC patients derived from a set of 1514 CRC cases. All 38 tumors had been tested in our laboratory for microsatellite instability (MSI), and Sanger sequencing had been used to screen for *MLH1* and *MSH2* mutations in MSI cases. Also, gastrointestinal polyposis had been diagnosed clinically and molecularly. Family histories were acquired from national registries. If inherited syndromes had not been diagnosed in routine diagnostic efforts ($n = 23$), normal tissue DNA was analyzed for mutations in a comprehensive set of high-penetrance genes (*MLH1, MSH2, MSH6, PMS2, APC, MUTYH, SMAD4, BMPR1A, LKB1*/*STK11,* and *PTEN*) by exome sequencing. ***Results.*** CRC predisposition syndromes were confirmed in 42% (16/38) of early-onset CRC patients. Hereditary nonpolyposis colorectal cancer was diagnosed in 12 (32%) patients, familial adenomatous polyposis in three (7.9%), and juvenile polyposis in one (2.6%) patient. Exome sequencing revealed one additional *MLH1* mutation. Over half of the patients had advanced cancers (Dukes C or D, 61%, 23/38). The majority of nonsyndromic patients had unaffected first-degree relatives and microsatellite-stable tumors. ***Conclusions.*** Microsatellite instability positivity or gastrointestinal polyposis characterized all patients with unambiguous highly penetrant germline mutations. In our series, exome sequencing produced little added value in diagnosing the underlying predisposition conditions.

Key Words: *age of onset, colorectal neoplasms, exome, genetic predisposition to disease, mutation*

Introduction

Early age of onset is a central characteristic of hereditary predisposition to cancer. Highly penetrant syndromes that predispose to colorectal cancer (CRC) at an early age include hereditary nonpolyposis colorectal cancer (HNPCC), familial adenomatous polyposis (FAP), juvenile polyposis (JP), Peutz-Jeghers syndrome (PJS) and *MUTYH*-associated polyposis (MAP). These syndromes can be classified into those with gastrointestinal polyposis (FAP, JP, PJS, and MAP), usually identified clinically, and those with a

Correspondence: Lauri A. Aaltonen, Department of Medical Genetics, Genome-Scale Biology Program, University of Helsinki, Biomedicum, P.O. Box 63 (Haartmaninkatu 8), FIN-00014 University of Helsinki, Finland. Tel: +358 9 19125595. Fax: +358 9 19125105. E-mail: lauri.aaltonen@helsinki.fi

(Received 21 December 2012; revised 21 January 2013; accepted 1 March 2013)

ISSN 0036-5521 print/ISSN 1502-7708 online © 2013 Informa Healthcare
DOI: 10.3109/00365521.2013.783102

more subtle, nonpolypotic phenotype (HNPCC, occasionally MAP) [1]. High penetrance is likely to explain less than 5% of all CRCs [2]. However, young patients are affected disproportionately.

The diagnosis of HNPCC was classically based on family history, but the introduction of immunohistochemical and molecular analyses has led to a considerable improvement in diagnostic accuracy [3,4]. Mismatch-repair (MMR) protein immunohistochemistry, testing for microsatellite instability (MSI), and Sanger sequencing of MMR genes are routine clinical practice in selected cases. Family history is formally assessed by the Amsterdam I, Amsterdam II, or Bethesda criteria [5–7].

It has been suggested that early-onset CRC tends to present at an advanced stage, probably reflecting a delay in presentation or diagnosis, or biological aggressiveness [8]. Also, several authors have observed adverse histopathological characteristics compared to older patients [9,10]. Other features associated with early age of onset are primary tumor location in the proximal colon, microsatellite instability, and family history of CRC [8,11,12].

Epidemiologically, the incidence of CRC in young adults, particularly that of rectal tumors, has increased in the last decades. This emphasizes environmental factors such as diet, obesity and low physical activity, acting independently or in concert with hereditary factors [11,13].

Colorectal cancer is one of the most preventable malignancies, and there is continuous debate on optimal screening strategies [14,15]. Screening programs have been devised for specific syndromes, but their effective implementation requires, among other factors, timely genetic diagnosis. The prevention of nonsyndromic early-onset CRC is a difficult task, since many of these young individuals might lack currently recognized risk factors such as personal or family history of colorectal neoplasia, or inflammatory bowel disease. Important unresolved questions concern the spectrum of germline mutations and the role of next-generation sequencing in their detection. Whole-genome and exome sequencing are rapidly becoming feasible options in the diagnosis of hereditary susceptibility, but little information is available on the usefulness of these approaches in early-onset CRC. The clinical application of next-generation sequencing would involve analyzing the protein-coding sequences of known CRC susceptibility genes in high-risk individuals. A conceptually similar, although technically different, approach was used by Sweet et al. [16] to analyze the genes *SMAD4*, *BMPR1A*, *STK11*/*LKB1*, and *ENG* in 49 patients with unexplained gastrointestinal polyposis. In this study, we aim at addressing these issues and provide an analysis of the molecular genetic, histopathological, clinical, and family history features of early-onset CRC, defined as age of onset less than 40 years.

Materials and methods

A previously described population-based material of normal and tumor tissue from 1042 CRC patients was used [4,17]. This material was collected between May 1994 and June 1998 in nine central hospitals covering southeastern Finland. The samples were studied for MSI, and in positive cases *MLH1* and *MSH2* Sanger sequencing was performed as described previously [4,17]. This series contributed 20 patients to the current study. Less systematic sample collection of unselected CRC cases was continued uninterruptedly in two of the nine central hospitals in 1998, and is still ongoing. An additional material of 472 CRCs and respective normal tissue samples was available from this additional series, and this series contributed 18 early-onset CRC cases. In these 18 tumors MSI testing was performed using the Bethesda panel of microsatellite markers (BAT25, BAT26, D5S346, D17S250, and D2S123), and MMR gene mutation data were obtained from diagnostic laboratories. Data on all first-degree relatives and their cancer diagnoses were acquired from official population registries and the Finnish Cancer Registry. Both the population registries and the Finnish Cancer Registry have almost complete coverage, and the Finnish Cancer Registry has been established in 1953. Medical records were used to obtain results from genetic testing and to investigate phenotypic features. This enabled us to take into account the clinical and molecular diagnoses of hereditary cancer readily available.

If a germline mutation had not been identified in the previous efforts (23 of 38 patients), germline protein-coding regions were sequenced by exome sequencing. SureSelect Human All Exon Kit v.1 (Agilent Technologies, Santa Clara, California) was used to capture exomic sequences, and prepared samples were sequenced by Illumina Genome Analyzer II (Illumina, San Diego, California) to obtain paired-end short read sequences. Read length was 80 base pairs and average coverage was 53.7. A comparative analysis tool (RikuRator, manuscript under preparation) was used to interpret exome sequencing data and call variants. Variants were filtered against 212 exome sequencing controls (68 in-house control exomes and 144 Finnish migraine patient exomes), and data from the 1000 Genomes Project (Phase 1 release, www.1000genomes.org), to exclude common polymorphisms. Subsequent analyses focused on the known high-penetrance CRC

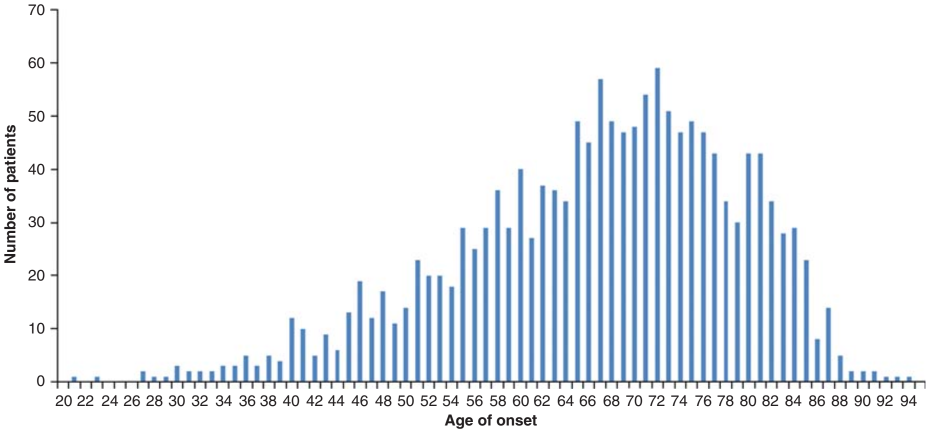


Figure 1. Age distribution of the entire CRC patient series (n = 1514).

genes, *MLH1*, *MSH2*, *MSH6*, *PMS2*, *APC*, *MUTYH*, *SMAD4*, *BMPR1A*, *LKB1*/*STK11*, and *PTEN*. Sanger sequencing was used to validate the relevant exome sequencing variants in these genes. All nonsilent variants in these genes with unknown or low (<5%) minor allele frequency were searched for in the InSiGHT mutation database (www.insight-group.org).

Results

The entire unselected CRC patient series comprised 1514 cases (Figure 1). Of these, 38 (2.6%) were diagnosed before the age of 40 years, ranging from 21 to 39 years. The median and mean ages of onset were 35 and 33.7, respectively. There were 18 females and 20 males.

Initially, all 38 young patients underwent testing for tumor MSI. If tumor tissue was microsatellite-unstable, germline mutations were screened for in the genes *MLH1* and *MSH2*. Diagnostic evaluation and clinical genetic testing revealed gastrointestinal polyposis syndromes and additional germline mutations. By these approaches, hereditary CRC was diagnosed in 15 of 38 patients. To examine the power of

Table I. Molecular, histopathological, clinical, and family history features of 38 early-onset CRC patients in an unselected Finnish series of 1514 CRC cases.

	Any CRC predisposition syndrome (n = 16)	HNPCC (n = 12)	No CRC predisposition syndrome diagnosed (n = 22)	MSI (n = 14)	All patients (n = 38)
Median age of onset	34.5	35	35.5	34.5	35
Mean age of onset	33.4	34.3	33.9	33.9	33.7
Gender					
Female	8 (50%)	7 (58%)	10 (45%)	7 (50%)	18 (47%)
Male	8 (50%)	5 (42%)	12 (55%)	7 (50%)	20 (53%)
MSI	12 (75%)	12 (100%)	2 (9.1%)	14 (100%)	14 (37%)
Syndromes					
HNPCC	12 (75%)	12 (100%)	0	12 (86%)	12 (32%)
FAP	3 (19%)	0	0	0	3 (7.9%)
JP	1 (6.3%)	0	0	0	1 (2.6%)
Primary location					
Proximal	5 (31%)	4 (33%)	4 (18%)	4 (29%)	9 (24%)
Distal	11 (69%)	8 (67%)	18 (82%)	10 (71%)	29 (76%)
Dukes stage					
A	3 (19%)	2 (17%)	3 (14%)	2 (14%)	6 (16%)
B	3 (19%)	3 (25%)	6 (27%)	4 (29%)	9 (24%)
C	8 (50%)	5 (42%)	9 (41%)	6 (43%)	17 (45%)
D	2 (13%)	2 (17%)	4 (18%)	2 (14%)	6 (16%)
First-degree relative with CRC	8 (50%)	7 (58%)	3 (14%)	7 (50%)	11 (29%)
Metachronous CRC	1 (6.3%)	1 (8.3%)	0	1 (7.1%)	1 (2.6%)
Synchronous CRC	1 (6.3%)	1 (8.3%)	0	1 (7.1%)	1 (2.6%)

Table II. Germline mutations in known CRC susceptibility genes in 16 CRC patients with age of onset less than 40 years. Mutation 1 is a Finnish founder mutation, a 3.5-kb genomic deletion comprising exon 16 of MLH1. The BMPR1A mutation g. 391C>G has been shown to result in skipping of exon 1 [30].

Patient	Age of onset	Gene	Mutation	Mutation type
s171	35	*MLH1*	c.454-1G>A	Splice site
s1108	38	*MLH1*	Mutation 1	Genomic deletion
s30	27	*MLH1*	c.1975C>T	Nonsense (R659X)
s17	27	*MLH1*	c.677+1G>T	Splice site
s124	34	*MLH1*	Mutation 1	Genomic deletion
c430	36	*MLH1*	Mutation 1	Genomic deletion
c615	39	*MLH1*	c.454-1G>A	Splice site
s49	33	*MSH2*	Deletion of exons 1-7	Genomic deletion
s108	38	*MSH2*	Deletion of exons 7-8	Genomic deletion
c138	39	*MSH2*	c.1807G>A	Missense (D603N)
c54	35	*MSH2*	c.1387-1G>T	Splice site
s95	31	*MSH6*	c.3013C>T	Nonsense (R1005X)
c520	29	FAP (clinical diagnosis)		
c231	34	*APC*	c.739C>T	Nonsense (Q247X)
s200	37	FAP (clinical diagnosis)		
c79	23	*BMPR1A*	g.391C>G	Splice site

exome sequencing in diagnosing additional cases of hereditary CRC in the series, the remaining 23 cases were analyzed utilizing this approach. The subsequent analysis focused on all the genes that have been implicated in high-penetrance CRC predisposition. This set comprised *MLH1*, *MSH2*, *MSH6*, *PMS2*, *APC*, *MUTYH*, *SMAD4*, *BMPR1A*, *LKB1/STK11*, and *PTEN*.

In exome sequencing, a splice-site mutation in *MLH1* (c.454-1G>A, the Finnish founder mutation 2) was found. The mutation was validated by PCR amplification and Sanger sequencing. The patient (s171, age of onset 35) had a microsatellite-unstable tumor, and thus it could have been possible to detect this mutation by MSI testing followed by *MLH1* Sanger sequencing. Variants leading to a truncated gene product were not found in other genes in this set. However, missense variants were found, most of them with dbSNP reference IDs (rs#). Missense variants with unknown or low minor allele frequencies (<5%) were searched for in the InSiGHT mutation database. All of these missense variants were found in the InSiGHT database, but none were classified as unambiguously pathogenic.

The main results are summarized in Tables I and II. A high proportion of the young cases display inherited CRC predisposition syndromes. Altogether, 42% (16/38) of the patients had clearly defined genetic susceptibility. The most frequent syndrome was HNPCC, which accounted for 12 cases (32%, 12/38). There were seven mutations in *MLH1* (18%, 7/38), four in *MSH2* (11%, 4/38) and one in *MSH6* (2.6%, 1/38). Incidentally, one patient (s49, age of onset 33) had HNPCC and multiple endocrine neoplasia type 1 (MEN-1).

Familial adenomatous polyposis was clinically diagnosed in three patients, and JP in one patient with a mutation in the gene *BMPR1A*. There was one case with X-linked agammaglobulinemia (XLA) and a microsatellite-unstable tumor. Exome sequencing did not reveal variation in the X-chromosomal *BTK* gene that most commonly causes XLA [18].

The majority of cancers had metastasized either to regional lymph nodes or to more distant tissues by the time of diagnosis (Dukes stage C or D, 61%, 23/38). The distribution of Dukes stages was as follows: stage A, 6 cases; stage B, 9 cases; stage C, 17 cases; and stage D, 6 cases. No difference in Dukes stage was seen between syndromic cases (Dukes C or D, 63% or 10/16) and nonsyndromic cases (Dukes C or D, 59% or 13/22). Tumors grades were available in the pathology reports of 35 of 38 tumors; 8.6% (3/35) were well differentiated (grade I), 74% (26/35) were moderately differentiated (grade II), and 17% (6/35) were poorly differentiated (grade III). Overall, 24% (9/38) of tumors were proximal to the splenic flexure, whereas 76% (29/38) were located in the distal colon or rectum. Based on pathology reports, inflammatory bowel disease was not present in any of the 38 young patients.

Of 38 tumors, 14 (37%) were microsatellite-unstable. MSI tumors correlated well with germline MMR gene mutations, since 86% (12/14) of these individuals had such mutations. Only 7.7% (2/26) of those without germline MMR gene mutations displayed MSI. Twenty-nine percent (4/14) of MSI tumors were located in the proximal colon, whereas 21% (5/24) of MSS tumors were proximal.

Complete family and personal cancer histories were obtained shortly after tumor resection by linking data from official population registries and the Finnish Cancer Registry. Eleven patients (29%, 11/38) were familial, i.e., at least one first-degree relative (parent, sibling, or child) was affected. Among patients with familial early-onset CRC, 73% (8/11) had a predisposing germline mutation (seven patients with MMR gene mutations and one patient with an APC mutation). In one patient with HNPCC (c138, age of onset 39), another CRC had been diagnosed at the age of 37 years (i.e., the patient had metachronous CRC), and synchronous CRC was diagnosed in the patient s124 (age of onset 34 years). In contrast, only 14% (3/22) of those without well-defined predisposition syndromes were familial, and none had synchronous or metachronous disease. One patient with FAP (c231) had two first-degree relatives with FAP-associated CRCs. However, the first-degree relatives of the other two patients with FAP were free of CRC.

Discussion

Our aims were to characterize the molecular, clinical, pathological and family history features of young CRC patients, and to evaluate the power of exome sequencing in diagnosing the underlying conditions. In an unselected Finnish series of 1514 CRC patients, only 2.6% (38 of 1514) of cases were early-onset, defined here as age of onset less than 40 years. In previously published data, estimates have ranged from 1.6% to 7.4% [8,11,19,20].

Mendelian disorders are notably common in early-onset CRC patients. In our series, the frequency of highly penetrant syndromes (42%, 16/38) was 10-fold higher compared to unselected CRC cases (<5%) [2]. HNPCC was the most prevalent form of predisposition, but there were also a smaller number of patients with gastrointestinal polyposis syndromes – three with FAP and one with JP – in line with the lower incidence rates of these syndromes.

Despite scrutinizing a comprehensive high-penetrance CRC gene set (*MLH1*, *MSH2*, *MSH6*, *PMS2*, *APC*, *MUTYH*, *SMAD4*, *BMPR1A*, *LKB1*/*STK11*, and *PTEN*), exome sequencing revealed only one additional mutation. However, it would have been possible to find this *MLH1* mutation by MSI testing followed by *MLH1* Sanger sequencing alone, since the tumor displayed MSI, but genetic testing data were not available from this patient. After screening for specific syndromes by established clinical and molecular approaches, the additional diagnostic yield of exome sequencing was low in our patient series, mainly producing missense variants of uncertain pathogenic significance. Exome sequencing has technical limitations in detecting insertions and deletions, copy number variants, mosaic mutations, and epigenetic alterations, which could at least partially explain the scarcity of detected mutations.

The fraction of MSI tumors was high (37%, 14/38). Germline MMR gene mutations were identified in the vast majority of young individuals with microsatellite-unstable tumors (86%, 12/14). On the contrary, only 7.7% (2/26) of those without germline MMR gene mutations had microsatellite-unstable tumors. Thus, tumor MSI was highly specific to germline MMR gene mutations in this patient population, and a careful search for MMR gene mutations is warranted in every early-onset patient with MSI CRC. Other authors have reported various frequencies of MSI and germline MMR gene mutations in patients under age 40 or 45 years, ranging from 4% to 31% for MSI and 5% to 19% for MMR gene mutations [9,21–24]. Single-center studies should be interpreted with caution, because there is an increased risk of distortion depending on local health policies and the expertise of different centers [25]. Our material was collected initially in nine central hospitals, but collection was limited to two central hospitals after 1042 samples, which is a potential source of bias, and could contribute to the relatively high number of patients with inherited syndromes. Indeed, the proportion of young cases was lower in the first, more systematic, phase of sample collection.

In line with a large body of evidence, a high proportion of advanced stage cancers were found (Dukes C or D, 61% or 23/38). We did not notice a correlation between tumor stage and CRC predisposition syndromes. The proportions of Dukes C or D stage cancers were 61%, 50%, 59% and 57% in syndromic, HNPCC, nonsyndromic, and MSI tumors, respectively. This highlights the need for early diagnosis, regardless of etiology. The primary location of MSI tumors showed only a slight predilection to the proximal colon compared to MSS tumors, since the fraction of proximal tumors was 29% (4/14) in MSI tumors, and 21% (5/24) in MSS tumors.

Our approach to family history was robust, since official population registries and the Finnish cancer registry were used to acquire accurate data shortly after surgical tumor resection to represent the family history at the time of diagnosis. Interview-based family histories are liable to omissions and misreporting, and often the reported cancers cannot be verified histopathologically. Familial early-onset CRC should raise high suspicion of inherited predisposition, as 73% (8/11) of these patients had predisposing mutations, seven with HNPCC and one with FAP. Among patients with HNPCC, affected first-degree relatives were present in 58% (7/12), and previous metachronous CRC in 8.3% (1/12). Similar histories were seen rarely in nonsyndromic individuals, as only

14% (3/22) had affected first-degree relatives, and none had metachronous disease.

Intriguingly, one patient (c543, age of onset 30 years) with XLA had developed microsatellite-unstable CRC. van der Meer et al. proposed that XLA causes a high risk of CRC [26]. This fits into the idea that immune deficiencies can disrupt cancer immune surveillance, leading to insufficient immune responses against tumor cells [27]. Microsatellite-unstable tumors are hypermutated, which gives rise to a large diversity of tumor antigens, and therefore an intact immune system might be essential in eliminating these neoplasms. The MSI pathway of colorectal tumorigenesis could be exceptionally effective in the context of immune deficiency, leading to the manifestation of CRC at an early age.

Two different hereditary tumor susceptibility syndromes, HNPCC and MEN-1 syndrome, had been diagnosed in the patient s49 (age of onset 33). Medical records were reviewed, and the most prominent clinical features were CRC, pituitary adenoma, hyperparathyroidism, and breast hypertrophy. This does not seem to indicate clear additive effects of the two gene defects.

In our series, syndromic early-onset CRC displayed autosomal dominant inheritance, typically there was personal or family history of CRC, and either gastrointestinal polyposis or microsatellite instability were present. In contrast, the etiology of nonsyndromic cases is obscure, first-degree relatives are mostly unaffected, polyposis is not a feature, and tumors are microsatellite-stable. Age of onset did not seem to differ between these two types. Possible etiologies for the nonsyndromic type include polygenic or recessive inheritance, rare variants with intermediate penetrance, and environmental factors, but further investigation is needed to discover risk factors in this poorly understood subset. A critical finding is that a genetic diagnosis could be feasible in as many as 40% of early-onset CRC patients. Testing for MSI and taking a family history of CRC can provide valuable clues that help to distinguish patients with HNPCC. Next-generation sequencing, including whole-genome and exome sequencing, is likely to become a key diagnostic technique in the near future [28,29]. Nevertheless, in our series, this approach provided little additional clues for genetic diagnosis as compared with testing for MSI and information on clinical features, in particular polyposis.

Acknowledgements

We are grateful to Sini Nieminen, Sirpa Soisalo, Inga-Lill Svedberg, Mairi Kuris, Sini Karjalainen, Kirsi Pylvänäinen, Tuula Lehtinen, and Iina Vuoristo for technical assistance. We acknowledge professor Aarno Palotie (FIMM and Wellcome Trust Sanger Institute, UK) and docent Maija Wessman (FIMM and Folhälsan Research Center, Helsinki) for providing Finnish, population-specific exome sequences for filtering.

Declaration of interest: The authors declare no competing interests. This work was supported by grants from the Academy of Finland (Finnish Center of Excellence Program 2012–2017), the Finnish Cancer Society, the Sigrid Juselius Foundation and SYSCOL (EU FP7 Collaborative Project).

References

[1] Lynch HT, de la Chapelle A. Hereditary colorectal cancer. N Engl J Med 2003;348:919–32.

[2] Aaltonen L, Johns L, Järvinen H, Mecklin JP, Houlston R. Explaining the familial colorectal cancer risk associated with mismatch repair (MMR)-deficient and MMR-stable tumors. Clin Cancer Res 2007;13:356–61.

[3] Thibodeau SN, French AJ, Roche PC, Cunningham JM, Tester DJ, Lindor NM, et al. Altered expression of hMSH2 and hMLH1 in tumors with microsatellite instability and genetic alterations in mismatch repair genes. Cancer Res 1996;56:4836–40.

[4] Aaltonen LA, Salovaara R, Kristo P, Canzian F, Hemminki A, Peltomäki P, et al. Incidence of hereditary nonpolyposis colorectal cancer and the feasibility of molecular screening for the disease. N Engl J Med 1998;338:1481–7.

[5] Vasen HF, Mecklin JP, Khan PM, Lynch HT. The International Collaborative Group on Hereditary Non-Polyposis Colorectal Cancer (ICG-HNPCC). Dis Colon Rectum 1991;34:424–5.

[6] Vasen HF, Watson P, Mecklin JP, Lynch HT. New clinical criteria for hereditary nonpolyposis colorectal cancer (HNPCC, Lynch syndrome) proposed by the International Collaborative group on HNPCC. Gastroenterology 1999; 116:1453–6.

[7] Umar A, Boland CR, Terdiman JP, Syngal S, de la Chapelle A, Rüschoff J, et al. Revised Bethesda Guidelines for hereditary nonpolyposis colorectal cancer (Lynch syndrome) and microsatellite instability. J Natl Cancer Inst 2004;96:261–8.

[8] Griffin PM, Liff JM, Greenberg RS, Clark WS. Adenocarcinomas of the colon and rectum in persons under 40 years old. A population-based study. Gastroenterology 1991;100:1033–40.

[9] Yantiss RK, Goodarzi M, Zhou XK, Rennert H, Pirog EC, Banner BF, et al. Clinical, pathologic, and molecular features of early-onset colorectal carcinoma. Am J Surg Pathol 2009; 33:572–82.

[10] Chang DT, Pai RK, Rybicki LA, Dimaio MA, Limaye M, Jayachandran P, et al. Clinicopathologic and molecular features of sporadic early-onset colorectal adenocarcinoma: an adenocarcinoma with frequent signet ring cell differentiation, rectal and sigmoid involvement, and adverse morphologic features. Mod Pathol 2012;25:1128–39.

[11] O'Connell JB, Maggard MA, Liu JH, Etzioni DA, Livingston EH, Ko CY. Rates of colon and rectal cancers are increasing in young adults. Am Surg 2003;69:866–72.

[12] Farrington SM, Lin-Goerke J, Ling J, Wang Y, Burczak JD, Robbins DJ, et al. Systematic analysis of hMSH2 and hMLH1 in young colon cancer patients and controls. Am J Hum Genet 1998;63:749–59.

[13] Siegel RL, Jemal A, Ward EM. Increase in incidence of colorectal cancer among young men and women in the United States. Cancer Epidemiol Biomarkers Prev 2009; 18:1695–8.

[14] Levin B, Lieberman DA, McFarland B, Andrews KS, Brooks D, Bond J, et al. Screening and surveillance for the early detection of colorectal cancer and adenomatous polyps, 2008: a joint guideline from the American Cancer Society, the US Multi-Society Task Force on Colorectal Cancer, and the American College of Radiology. Gastroenterology 2008;134:1570–95.

[15] Rex DK, Johnson DA, Anderson JC, Schoenfeld PS, Burke CA, Inadomi JM. American College of Gastroenterology guidelines for colorectal cancer screening 2009 [corrected]. Am J Gastroenterol 2009;104:739–50.

[16] Sweet K, Willis J, Zhou XP, Gallione C, Sawada T, Alhopuro P, et al. Molecular classification of patients with unexplained hamartomatous and hyperplastic polyposis. JAMA 2005;294:2465–73.

[17] Salovaara R, Loukola A, Kristo P, Kääriäinen H, Ahtola H, Eskelinen M, et al. Population-based molecular detection of hereditary nonpolyposis colorectal cancer. J Clin Oncol 2000;18:2193–200.

[18] Vetrie D, Vorechovský I, Sideras P, Holland J, Davies A, Flinter F, et al. The gene involved in X-linked agammaglobulinaemia is a member of the src family of protein-tyrosine kinases. Nature 1993;361:226–33.

[19] Adkins RB, DeLozier JB, McKnight WG, Waterhouse G. Carcinoma of the colon in patients 35 years of age and younger. Am Surg 1987;53:141–5.

[20] Adloff M, Arnaud JP, Schloegel M, Thibaud D, Bergamaschi R. Colorectal cancer in patients under 40 years of age. Dis Colon Rectum 1986;29:322–5.

[21] Liang JT, Huang KC, Cheng AL, Jeng YM, Wu MS, Wang SM. Clinicopathological and molecular biological features of colorectal cancer in patients less than 40 years of age. Br J Surg 2003;90:205–14.

[22] Perea J, Alvaro E, Rodríguez Y, Gravalos C, Sánchez-Tomé E, Rivera B, et al. Approach to early-onset colorectal cancer: clinicopathological, familial, molecular and immunohistochemical characteristics. World J Gastroenterol 2010; 16:3697–703.

[23] Losi L, Di Gregorio C, Pedroni M, Ponti G, Roncucci L, Scarselli A, et al. Molecular genetic alterations and clinical features in early-onset colorectal carcinomas and their role for the recognition of hereditary cancer syndromes. Am J Gastroenterol 2005;100:2280–7.

[24] Southey MC, Jenkins MA, Mead L, Whitty J, Trivett M, Tesoriero AA, et al. Use of molecular tumor characteristics to prioritize mismatch repair gene testing in early-onset colorectal cancer. J Clin Oncol 2005;23:6524–32.

[25] Terdiman JP, Levin TR, Allen BA, Gum JR, Fishbach A, Conrad PG, et al. Hereditary nonpolyposis colorectal cancer in young colorectal cancer patients: high-risk clinic versus population-based registry. Gastroenterology 2002;122:940–7.

[26] van der Meer JW, Weening RS, Schellekens PT, van Munster IP, Nagengast FM. Colorectal cancer in patients with X-linked agammaglobulinaemia. Lancet 1993;341:1439–40.

[27] Dunn GP, Bruce AT, Ikeda H, Old LJ, Schreiber RD. Cancer immunoediting: from immunosurveillance to tumor escape. Nat Immunol 2002;3:991–8.

[28] Ashley EA, Butte AJ, Wheeler MT, Chen R, Klein TE, Dewey FE, et al. Clinical assessment incorporating a personal genome. Lancet 2010;375:1525–35.

[29] Choi M, Scholl UI, Ji W, Liu T, Tikhonova IR, Zumbo P, et al. Genetic diagnosis by whole exome capture and massively parallel DNA sequencing. Proc Natl Acad Sci USA 2009;106:19096–101.

[30] Zhou XP, Woodford-Richens K, Lehtonen R, Kurose K, Aldred M, Hampel H, et al. Germline mutations in BMPR1A/ALK3 cause a subset of cases of juvenile polyposis syndrome and of Cowden and Bannayan-Riley-Ruvalcaba syndromes. Am J Hum Genet 2001;69:704–11.

II

ELSEVIER

Cancer Genetics 208 (2015) 35–40

Cancer Genetics

Systematic search for rare variants in Finnish early-onset colorectal cancer patients

Tomas Tanskanen [a], Alexandra E. Gylfe [a], Riku Katainen [a], Minna Taipale [b,e], Laura Renkonen-Sinisalo [a,c], Heikki Järvinen [c], Jukka-Pekka Mecklin [d], Jan Böhm [d], Outi Kilpivaara [a], Esa Pitkänen [a], Kimmo Palin [a], Pia Vahteristo [a], Sari Tuupanen [a], Lauri A. Aaltonen [a,*]

[a] *Department of Medical Genetics, Genome-Scale Biology Program, University of Helsinki, Biomedicum, Helsinki, Finland;* [b] *Genome-Scale Biology Program, Institute of Biomedicine, University of Helsinki, Biomedicum, Helsinki, Finland;* [c] *Department of Surgery, Helsinki University Central Hospital, Helsinki, Finland;* [d] *Department of Surgery, Jyväskylä Central Hospital, University of Eastern Finland, Jyväskylä, Finland;* [e] *Science for Life Center, Department of Biosciences and Nutrition, Karolinska Institute, Huddinge, Sweden*

The heritability of colorectal cancer (CRC) is incompletely understood, and the contribution of undiscovered rare variants may be important. In search of rare disease-causing variants, we exome sequenced 22 CRC patients who were diagnosed before the age of 40 years. Exome sequencing data from 95 familial CRC patients were available as a validation set. Cases with known CRC syndromes were excluded. All patients were from Finland, a country known for its genetically homogenous population. We searched for rare nonsynonymous variants with allele frequencies below 0.1% in 3,374 Finnish and 58,112 non-Finnish controls. In addition, homozygous and compound heterozygous variants were studied. No genes with rare loss-of-function variants were present in more than one early-onset CRC patient. Three genes (*ADAMTS4*, *CYTL1*, and *SYNE1*) harbored rare loss-of-function variants in both early-onset and familial CRC cases. Five genes with homozygous variants in early-onset CRC cases were found (*MCTP2*, *ARHGAP12*, *ATM*, *DONSON*, and *ROS1*), including one gene (*MCTP2*) with a homozygous splice site variant. All discovered homozygous variants were exclusive to one early-onset CRC case. Independent replication is required to associate the discovered variants with CRC. These findings, together with a lack of family history in 19 of 22 (86%) early-onset patients, suggest genetic heterogeneity in unexplained early-onset CRC patients, thus emphasizing the requirement for large sample sizes and careful study designs to elucidate the role of rare variants in CRC susceptibility.

Keywords Genetic predisposition to disease, colorectal neoplasms, age of onset, exome sequencing

Colorectal cancer (CRC) accounts for 10% of new cancers worldwide (GLOBOCAN Project, http://globocan.iarc.fr), and its incidence rises rapidly after 45 years of age. The lifetime risk of CRC is approximately 5%, whereas the risk of developing CRC before the age of 40 years is only 0.08% (SEER database, http://seer.cancer.gov). The etiology of most CRCs is complex and multifactorial, involving interplay between multiple genetic and environmental factors. Inherited factors contribute to CRC risk considerably, but a significant fraction of heritability in CRC patients remains incompletely understood (1).

An estimated 5% of CRC patients, including many of those with early-onset disease or multiple affected family members, are highly predisposed to CRC because of rare single-gene defects in *MLH1*, *MSH2*, *MSH6*, *PMS2*, *APC*, *MUTYH*, *SMAD4*, *BMPR1A*, *STK11*/*LKB1*, or *POLE* (2). Because of large effect sizes, family-based linkage analysis was instrumental in mapping these genes. On the other hand, several low-penetrance CRC susceptibility loci have been discovered through genome-wide association studies

Received September 19, 2014; received in revised form December 12, 2014; accepted December 17, 2014.

* Corresponding author.
E-mail address: lauri.aaltonen@helsinki.fi

2210-7762/$ - see front matter © 2015 Elsevier Inc. All rights reserved.
http://dx.doi.org/10.1016/j.cancergen.2014.12.004

(3), but the allelic architectures and causative variants underlying these associations are mostly undefined.

Next generation sequencing (NGS) has uncovered patterns of human genetic variation in unprecedented detail. Because exome sequencing captures a substantial part of functional and disease-causing genomic variation, it is a promising approach to deciphering the role of rare variants in complex disease predisposition (4). Interest in the pathogenic potential of rare variants has emerged from evolutionary theory (5), as well as the fact that much of the heritability of complex diseases remains unexplained (6).

Despite the potential of NGS in the identification of complex trait genes, success has been limited. Part of the reason is genetic heterogeneity, which increases the sample size required for the identification of culprit genes. In this regard, population isolates could offer unique advantages. The population history of Finland has been characterized in detail. Population bottlenecks and genetic founder effects, as well as geographic isolation, have shaped the genetic structure of the population, leading to reduced genetic heterogeneity (7).

Early age of onset is a key feature of hereditary susceptibility to CRC and other common cancers (8), and early-onset CRC patients might be enriched for undiscovered susceptibility variants. In this study, we exome sequenced a discovery set of 22 unselected Finnish CRC patients who were diagnosed before the age of 40 years, and we used exome sequencing data from 95 Finnish familial CRC patients as a validation set. None of the patients displayed known predisposition syndromes that could account for early age of onset or a positive family history. To identify new potential CRC susceptibility genes, we analyzed rare nonsynonymous variants, and considered both dominant and recessive modes of inheritance.

Materials and methods

Samples

We studied a discovery set of 22 nonsyndromic early-onset CRC cases diagnosed before the age of 40 years, and we used 95 familial CRC cases (with at least one affected first-degree relative) as a validation set. Both sample sets were published previously (9,10) and were derived from a series of 1,514 unselected CRC patients, which was collected in nine central hospitals in southern and eastern Finland between May 1994 and June 1998 and in two of these hospitals from year 1998 to present (9,10). The population-based phase of sample collection in nine hospitals contributed 1,042 CRC patients, and 472 additional CRC patient samples were collected in two hospitals after June 1998. All patients gave informed consent to genetic studies on tumor susceptibility, and the study was approved by the appropriate ethics review board.

Both normal and tumor DNA samples were available from each patient. All tumors had been tested for microsatellite instability (MSI), and known CRC susceptibility syndromes had been diagnosed clinically or molecularly. Data on first-degree relatives and their cancer diagnoses had been acquired from official population registries and the Finnish Cancer Registry (9,10). Of 1,514 CRC patients, 38 (2.5%) had been diagnosed before the age of 40 years. Of 38 early-onset CRC patients, 16 (42%) had known genetic CRC susceptibility syndromes, including hereditary non-polyposis colon cancer (12 of 38, 32%), familial adenomatous polyposis (3 of 38, 7.9%), and juvenile polyposis (1 of 38, 2.6%). Based on pathology reports, there was no evidence of inflammatory bowel disease in any of the 38 early-onset patients. Of the 22 early-onset CRC patients with unknown etiology (Table 1), 10 were female (45%), 12 were male (55%), 2 displayed MSI (9.1%), 18 had cancers of the distal colon or rectum (82%), 13 presented with advanced-stage cancer (Dukes stage C or D, 59%), and 3 had a family history of CRC (14%). Median and mean ages of onset were 35.5 and 33.9 years, respectively, ranging from 21−39 years. Germline DNA samples of these 22 nonsyndromic CRC patients were exome sequenced.

Exome sequencing

Exome sequences were captured with the SureSelect Human All Exon Kit v.1 (Agilent Technologies, Santa Clara, CA). Paired-end 75 base pair reads were obtained with an Illumina HiSeq 2000 (Illumina, San Diego, CA). Exome sequencing data quality was confirmed with FastQC (http://www.bioinformatics.bbsrc.ac.uk/projects/fastqc). Reads were mapped to the human reference genome GRCh37 with the Burrows-Wheeler Aligner, v.0.5.9-r16. The Picard MarkDuplicates tool (http://broadinstitute.github.io/picard/) was used to remove duplicate reads. Reads were realigned locally with the Genome Analysis Toolkit IndelRealigner, and single-nucleotide variants and indels were called with the Genome Analysis Toolkit UnifiedGenotyper, v.2.2-16-g9f648cb (https://www.broadinstitute.org/gatk/). Average coverage was 54, and 87% of the targeted regions were covered by more than 10 reads. An in-house developed comparative analysis tool (RikuRator, unpublished) was used to determine allele frequencies in control exomes and to compare variant calls between CRC cases. When relevant, sequencing reads were manually inspected to exclude false-positive variant calls. Loss-of-function (LoF) variants were annotated with the LOFTEE (Loss-Of-Function Transcript Effect Estimator, https://github.com/konradjk/loftee). Accordingly, we excluded ancestral LoF alleles, LoF variants located in the last 5% of the coding region, splice site variants in small introns (<15 base pairs), and LoF variants surrounded by non-canonical splice sites. Functional effects of missense variants were predicted with PolyPhen-2 and SIFT, using the Ensembl Variant Effect Predictor (http://www.ensembl.org).

Exome sequencing controls

Allele frequencies of all variants were determined in 3,374 Finnish and 58,112 non-Finnish control exomes that were publicly available in the Exome Aggregation Consortium (ExAC) database (http://exac.broadinstitute.org). These individuals had been sequenced in various medical and population genetic studies.

Table 1 Clinical characteristics of the 22 exome sequenced early-onset CRC patients

Patient no.	Age of onset	Dukes stage	Primary location	MSI status	CRC in first-degree relatives
c206	36	C	Distal	MSS	Sibling, age 53
c270	39	D	Proximal	MSS	Parent, age 76
c386	30	C	Distal	MSS	Parent, age 56
c414	36	C	Proximal	MSS	—
c543	30	B	Distal	MSI	—
c592	30	A	Distal	MSS	—
c690	36	B	Distal	MSS	—
c768	32	C	Distal	MSI	—
c837	39	B	Distal	MSS	—
c907	34	A	Distal	MSS	—
c938	38	C	Distal	MSS	—
c1055	38	C	Distal	MSS	—
c1066	31	C	Distal	MSS	—
s154	38	A	Distal	MSS	—
s160	33	C	Distal	MSS	—
s281	37	B	Proximal	MSS	—
s907	21	D	Distal	MSS	—
s1137	31	D	Proximal	MSS	—
s1151	36	D	Distal	MSS	—
s1152	28	D	Distal	MSS	—
s1165	37	C	Distal	MSS	—
s1167	35	B	Distal	MSS	—

Sanger sequencing and genotyping

Relevant exome sequencing variant calls were verified by Sanger sequencing. The AmpliTaq Gold enzyme (Applied Biosystems, Foster City, CA) was used in PCR reactions, and PCR products were purified with the ExoSAP-IT PCR purification kit (USB Corporation, Cleveland, OH). Big Dye Terminator v3.1 chemistry (Applied Biosystems) was used in the DNA sequencing, and capillary electrophoresis was performed on an Applied Biosystems 3730xl DNA analyzer at the Institute for Molecular Medicine Finland (FIMM). All novel variants were submitted to dbSNP (http://www.ncbi.nlm.nih.gov/SNP). Allelic imbalance was scored by comparing Sanger sequencing peak heights in normal and tumor DNA samples as described previously (11).

Results

We exome sequenced 22 unselected, nonsyndromic CRC patients with disease onset before the age of 40 years (Table 1). Median and mean ages of onset were 35.5 and 33.9 years, respectively. Most patients (19 of 22, 86%) had no first-degree relatives with CRC.

In the exome sequencing, a total of 856,325 protein-coding variants were called, 652,626 of which were synonymous and 203,699 nonsynonymous. Of the nonsynonymous variant calls, 6,120 were classified as protein-truncating (nonsense, splice site, and frameshift variants), and the numbers of nonsense, splice site, and frameshift variants were 1,920, 1,613, and 2,587, respectively. Allele frequencies were determined in two exome control sets, the first consisting of 3,374 Finnish controls and the second of 58,112 non-Finnish controls. We hypothesized that early-onset CRC patients may carry rare, dominant CRC predisposing variants, and thus we searched for rare variants with allele frequencies of $<0.1\%$ in both control sets. The possibility that early-onset CRC patients carry recessive susceptibility variants was also considered. To this end, we excluded variants that were homozygous in any of the 3,374 Finnish or 58,112 non-Finnish controls. Finally, we shortlisted two sets of genes: 1) genes with heterozygous LoF or missense variants (Table 2), and 2) genes with homozygous nonsynonymous variants (Table 3).

Initially, protein-truncating variants were analyzed (Table 2). We found no genes with rare LoF variants in two or more early-onset CRC cases. Next, we identified all genes with a rare LoF variant in a single early-onset CRC case. Data from 95 familial CRC cases were then analyzed for rare LoF variants in these genes, and three genes harbored potential LoF variants in both sample sets (*ADAMTS4*, *CYTL1*, and *SYNE1*). All three LoF variants found in familial CRC patients were exclusive to one family. Familial CRC patients with LoF variants in *ADAMTS4*, *CYTL1*, and *SYNE1* had been diagnosed at ages 75, 86, and 84, respectively.

Next, we searched for rare missense variants shared by two or more early-onset CRC patients (Table 2). *ACSL5* p.Pro71Leu and *INTS5* p.Pro922Leu were both found in two early-onset CRC patients. Neither of these missense variants was found in 95 familial CRC cases. PolyPhen-2 and SIFT scores are shown in the Supplementary Table. *INTS5* p.Pro922Leu was classified as damaging by both prediction methods.

To see whether loss of the wild-type allele of heterozygous variants had occurred in the respective, paired cancer samples, allelic imbalance was analyzed by Sanger sequencing. Sanger sequencing did not suggest allelic imbalance for any of the variants.

Table 2 Rare LoF and missense variants in early-onset CRC patients

			CRC cases		Finnish controls		Non-Finnish controls	
Gene	cDNA	Amino acid	Young[a]	Familial[b]	Frequency[b]	Coverage	Frequency[b]	Coverage
ADAMTS4	c.1618delG	Frameshift	1 of 22	1 of 95	2 of 3,374	100.0%	0 of 58,112	99.4%
CYTL1	c.327+2T>A	Splice site	1 of 22	1 of 95	0 of 3,374	100.0%	19 of 58,112	99.9%
SYNE1	c.1941dupT	Frameshift	1 of 22	0 of 95	0 of 3,374	95.7%	2 of 58,112	90.6%
SYNE1	c.5568delC	Frameshift	0 of 22	1 of 95	0 of 3,374	100.0%	0 of 58,112	99.8%
ACSL5	c.212C>T	p.Pro71Leu	2 of 22	0 of 95	5 of 3,374	99.9%	10 of 58,112	100.0%
INTS5	c.2765C>T	p.Pro922Leu	2 of 22	0 of 95	0 of 3,374	99.9%	2 of 58,112	99.7%

[a] Age of onset <40 years.

[b] Familial CRC cases had at least one affected first-degree relative.

Finally, we investigated possible recessive inheritance. In exome sequencing, 69,503 homozygous nonsynonymous variants were called. All variants found homozygously in any of 61,486 exome controls were excluded. This resulted in the identification of four genes with homozygous missense variants (*ARHGAP12*, *ATM*, *DONSON*, and *ROS1*) and one with a homozygous splice-site variant (*MCTP2* c.1488+1G>C). Allele frequencies of all homozygous variants were <1% in 3,374 Finnish controls. *MCTP2* c.1488+1G>C was homozygous in one early-onset CRC case and heterozygous in five familial CRC cases. Using the same filtering strategy as for homozygous variants, we did not find genes with compound heterozygous LoF variants. The analysis of compound heterozygous missense variants was complicated by haplotype-phase ambiguity and lack of individual-level control data, and was not pursued further.

Discussion

Understanding the genetic architecture of CRC susceptibility is of considerable interest clinically, since morbidity and mortality from the disease can be effectively reduced in patients known to be at high genetic risk (12). Early age of onset, along with familial aggregation, is a hallmark of inherited cancer susceptibility. In the general population, the median age for the diagnosis of CRC is 69 years, and only approximately 3% of CRCs occur in patients <40 years old (SEER database, http://seer.cancer.gov). Undiscovered disease-predisposing variants could be several-fold enriched in such extreme-phenotype populations, making it possible to identify rare causal variants even in relatively small sets of carefully selected cases (13), although it is challenging to demonstrate statistically significant enrichment in a genome-wide context. To test this approach, we searched for potentially disease-related rare variants in 22 exome sequenced CRC patients diagnosed before the age of 40 years, leveraging the homogenous population structure of Finland and a validation set of 95 familial CRC patients.

The majority (19 of 22, 86%) of the early-onset CRC patients had no first-degree relatives with CRC. At least to some degree, this argues against highly penetrant, dominant inheritance in many of the cases. More plausible genetic models include polygenic inheritance, gene-environment interactions, de novo mutations, and recessive inheritance. Also, coincidence is likely to play a role in a subset of cases.

We studied heterozygous variants with allele frequencies <0.1%. This relatively strict threshold was chosen to increase the proportion of functionally significant variants, although investigating more common low-frequency variants would also be relevant. In general, setting a low allele frequency threshold could allow efficient prioritization of variants, but this also results in higher estimated relative risks for variants found at notable frequencies among cases, making it unlikely to discover pathogenic variants with low penetrance.

Undiscovered LoF (nonsense, splice site, or frameshift) and missense variants are likely to contribute to CRC susceptibility, since they often cause direct alterations in gene function and are associated with a multitude of clinical conditions. It appears that healthy humans typically carry approximately 100 LoF variants, many of which are mildly deleterious and often found at low frequencies (14). The large number of deficient alleles per person complicates the functional interpretation of individual variants. We used LOFTEE as part of the filtering strategy to exclude predicted LoF variants that are unlikely to abolish gene function. To quantitatively evaluate how well human genes tolerate functional variation, Petrovski et al. estimated intolerance

Table 3 Homozygous variants in early-onset CRC patients

			CRC cases		Finnish controls		Non-Finnish controls	
Gene	cDNA	Amino acid	Young[a]	Familial[b]	Frequency[b]	Coverage	Frequency[b]	Coverage
MCTP2	c.1488+1G>C	Splice site	1 of 22	5 of 95	52 of 3,374	98.5%	71 of 58,112	98.3%
ARHGAP12	c.596G>C	p.Cys199Ser	1 of 22	0 of 95	14 of 3,374	99.9%	16 of 58,112	99.8%
ATM	c.998C>T	p.Ser333Phe	1 of 22	1 of 95	31 of 3,374	99.7%	125 of 58,112	99.6%
DONSON	c.1411G>A	p.Glu471Lys	1 of 22	2 of 95	37 of 3,374	100.0%	122 of 58,112	100.0%
ROS1	c.1108T>C	p.Ser370Pro	1 of 22	0 of 95	5 of 3,374	99.4%	198 of 58,112	99.0%

[a] Age of onset <40 years.

[b] Frequency of heterozygous variant carriers. Familial CRC cases had at least one affected first-degree relative.

scores (Residual Variation Intolerance Score, RVIS) for 16,956 human genes (15). Seven of the ten genes found in this study had RVIS values between the 25th and 75th percentiles (Supplementary Table). However, *INTS5* (RVIS −0.972, 8.95 percentile) and *ARHGAP12* (RVIS −0.819, 11.9 percentile) were estimated to be intolerant to functional variation, whereas *ATM* seemed remarkably tolerant (RVIS 1.53, 95.5 percentile).

Three genes (*ADAMTS4*, *CYTL1*, and *SYNE1*) harbored rare LoF variants in one early-onset CRC patient and one familial CRC patient. Instead of undertaking joint analysis, we used the 95 Finnish familial CRC cases as a validation set. This was because rare LoF variants had been previously studied in the 95 familial CRC cases, resulting in the identification of 11 candidate CRC susceptibility genes (16). The vast majority (85 of 95, 89%) of familial CRC cases had only one affected first-degree relative, compatible with a multifactorial etiology. In contrast to early-onset CRC cases, familial CRC patients who carried LoF variants in *ADAMTS4*, *CYTL1*, and *SYNE1* had relatively late ages of onset (75, 86, and 84, respectively). None of the variants displayed an allelic imbalance in tumor tissue.

Most of the genes found in this study have not been previously implicated in human cancer. Exceptions were *ATM* and *ROS1*, which have well-established roles in tumorigenesis (The Cancer Gene Census, http://cancer.sanger.ac.uk/cancergenome/projects/census/). However, the large variety of rare missense variants in *ATM* is a known complication in medical genetic studies (17), and since *ROS1* is a dominant oncogene (18), it would not be clear-cut to hypothesize that the homozygous variant p.Ser370Pro in *ROS1* predisposes to cancer, although this cannot be excluded. In *MCTP2*, we found a homozygous splice site variant in one early-onset CRC patient, possibly indicating complete gene inactivation, which is relatively rare in human genomes (19). Heterozygous *MCTP2* c.1488+1G>C was also found in 5 of the 95 familial CRC patients. In one study, it was suggested that *MCTP2* is a dosage-dependent gene required for cardiac development (20), which may contradict the interpretation that homozygous c.1488+1G>C causes complete inactivation of the gene.

Recently, there has been intense interest in using exome sequencing to investigate the genetic underpinnings of common diseases. Optimal study design depends heavily on the underlying genetic architecture, which can only be determined experimentally. In this analysis, we provide empirical data suggesting genetic heterogeneity in unexplained early-onset CRC patients in the Finnish founder population. The main findings supporting this conclusion are: 1) lack of rare, shared LoF variants between young CRC patients, 2) exclusiveness of discovered homozygous variants to single early-onset CRC cases, and 3) inconspicuous registry-based family histories, revealing affected first-degree relatives in only 3 of 22 early-onset CRC patients. Although the sample size was small, we attempted to enrich for causal variants by extreme-phenotype sampling in an isolated population, and protein-coding genes were systematically analyzed to prioritize variants that are biologically most plausible. The variants found in this study may be of interest in forthcoming genetic studies on multifactorial CRC susceptibility, but unless independent validation provides clear statistical evidence to support them, a high degree of caution must be taken to avoid biased interpretation. Based on these observations and recent developments in statistical methods (21), it would be compelling to test for an association between rare variants and complex CRC exome-wide. This would require large sample sizes, but would also produce data that is amenable to meta-analysis.

Acknowledgments

We are grateful to Sini Nieminen, Sirpa Soisalo, Inga-Lill Svedberg, Mairi Kuris, Sini Karjalainen, Kirsi Pylvänäinen, Tuula Lehtinen, and Iina Vuoristo for their technical assistance. This work was supported by grants from the Academy of Finland (Finnish Center of Excellence Program, 250345), the European Research Council (268648-NGG), SYSCOL (an EU FP7 Collaborative Project, 268648-NGG), the Cancer Society of Finland, and the Sigrid Juselius Foundation. We thank the Finnish Medical Foundation and the Finnish Medical Society Duodecim for a personal grant to Tomas Tanskanen, and University of Helsinki for a post-doctoral grant to Sari Tuupanen.

Supplementary data

Supplementary data related to this article can be found at http://dx.doi.org/10.1016/j.cancergen.2014.12.004

References

1. Lichtenstein P, Holm NV, Verkasalo PK, et al. Environmental and heritable factors in the causation of cancer—analyses of cohorts of twins from Sweden, Denmark, and Finland. N Engl J Med 2000;343:78—85.
2. Kilpivaara O, Aaltonen LA. Diagnostic cancer genome sequencing and the contribution of germline variants. Science 2013;339:1559—1562.
3. Lubbe SJ, Di Bernardo MC, Broderick P, et al. Comprehensive evaluation of the impact of 14 genetic variants on colorectal cancer phenotype and risk. Am J Epidemiol 2012;175:1—10.
4. Kiezun A, Garimella K, Do R, et al. Exome sequencing and the genetic basis of complex traits. Nat Genet 2012;44:623—630.
5. Pritchard JK, Cox NJ. The allelic architecture of human disease genes: common disease-common variant … or not? Hum Mol Genet 2002;11:2417—2423.
6. Manolio TA, Collins FS, Cox NJ, et al. Finding the missing heritability of complex diseases. Nature 2009;461:747—753.
7. Peltonen L, Palotie A, Lange K. Use of population isolates for mapping complex traits. Nat Rev Genet 2000;1:182—190.
8. Foulkes WD. Inherited susceptibility to common cancers. N Engl J Med 2008;359:2143—2153.
9. Aaltonen LA, Salovaara R, Kristo P, et al. Incidence of hereditary nonpolyposis colorectal cancer and the feasibility of molecular screening for the disease. N Engl J Med 1998;338:1481—1487.
10. Salovaara R, Loukola A, Kristo P, et al. Population-based molecular detection of hereditary nonpolyposis colorectal cancer. J Clin Oncol 2000;18:2193—2200.
11. Tuupanen S, Niittymäki I, Nousiainen K, et al. Allelic imbalance at rs6983267 suggests selection of the risk allele in somatic colorectal tumor evolution. Cancer Res 2008;68:14—17.
12. Lieberman DA. Clinical practice. Screening for colorectal cancer. N Engl J Med 2009;361:1179—1187.

13. Cirulli ET, Goldstein DB. Uncovering the roles of rare variants in common disease through whole-genome sequencing. Nat Rev Genet 2010;11:415–425.
14. MacArthur DG, Balasubramanian S, Frankish A, et al. A systematic survey of loss-of-function variants in human protein-coding genes. Science 2012;335:823–828.
15. Petrovski S, Wang Q, Heinzen EL, et al. Genic intolerance to functional variation and the interpretation of personal genomes. PLoS Genet 2013;9:e1003709.
16. Gylfe AE, Katainen R, Kondelin J, et al. Eleven candidate susceptibility genes for common familial colorectal cancer. PLoS Genet 2013;9:e1003876.
17. Scott SP, Bendix R, Chen P, et al. Missense mutations but not allelic variants alter the function of ATM by dominant interference in patients with breast cancer. Proc Natl Acad Sci U S A 2002;99:925–930.
18. Rikova K, Guo A, Zeng Q, et al. Global survey of phosphotyrosine signaling identifies oncogenic kinases in lung cancer. Cell 2007;131:1190–1203.
19. Lim ET, Raychaudhuri S, Sanders SJ, et al. Rare complete knockouts in humans: population distribution and significant role in autism spectrum disorders. Neuron 2013;77:235–242.
20. Lalani SR, Ware SM, Wang X, et al. MCTP2 is a dosage-sensitive gene required for cardiac outflow tract development. Hum Mol Genet 2013;22:4339–4348.
21. Lee S, Abecasis GR, Boehnke M, et al. Rare-variant association analysis: study designs and statistical tests. Am J Hum Genet 2014;95:5–23.

III

uicc
global cancer control

IJC
International Journal of Cancer

Short Report

Genome-wide association study and meta-analysis in Northern European populations replicate multiple colorectal cancer risk loci

Tomas Tanskanen [1,2], Linda van den Berg[1,2], Niko Välimäki[1,2], Mervi Aavikko[1,2], Eivind Ness-Jensen[3,4,5,6], Kristian Hveem[3,4], Yvonne Wettergren[7], Elinor Bexe Lindskog[7], Neeme Tõnisson[8], Andres Metspalu[8], Kaisa Silander[9], Giulia Orlando[10], Philip J. Law [10], Sari Tuupanen[1,2], Alexandra E. Gylfe[1,2], Ulrika A. Hänninen[1,2], Tatiana Cajuso[1,2], Johanna Kondelin[1,2], Antti-Pekka Sarin[11], Eero Pukkala[12,13], Pekka Jousilahti[14], Veikko Salomaa[14], Samuli Ripatti[11], Aarno Palotie[11,15,16,17], Heikki Järvinen[18], Laura Renkonen-Sinisalo[18], Anna Lepistö[18], Jan Böhm[19], Jukka-Pekka Mecklin[20], Nada A. Al-Tassan[21], Claire Palles[22], Lynn Martin[23], Ella Barclay[22], Albert Tenesa[24,25], Susan M. Farrington[24], Maria N. Timofeeva[24], Brian F. Meyer[21], Salma M. Wakil[21], Harry Campbell[26], Christopher G. Smith[27], Shelley Idziaszczyk[27], Tim S. Maughan[28], Richard Kaplan[29], Rachel Kerr[30], David Kerr[31], Daniel D. Buchanan [32,33], Aung K. Win [33], John Hopper[33], Mark A. Jenkins[33], Polly A. Newcomb[34], Steve Gallinger[35], David Conti[36], Fredrick R. Schumacher[37], Graham Casey[38], Jeremy P. Cheadle[27], Malcolm G. Dunlop[24], Ian P. Tomlinson[23], Richard S. Houlston[10], Kimmo Palin[1,2] and Lauri A. Aaltonen[1,2]

[1] Department of Medical and Clinical Genetics, Medicum, University of Helsinki, Helsinki, Finland
[2] Genome-Scale Biology Research Program, Research Programs Unit, University of Helsinki, Helsinki, Finland
[3] HUNT Research Centre, Department of Public Health, Norwegian University of Science and Technology (NTNU), Levanger, Norway
[4] K.G. Jebsen Center for Genetic Epidemiology, Department of Public Health, Norwegian University of Science and Technology (NTNU), Trondheim, Norway
[5] Department of Molecular Medicine and Surgery, Karolinska Institutet, Karolinska University Hospital, Stockholm, Sweden
[6] Department of Medicine, Levanger Hospital, Nord-Trøndelag Hospital Trust, Levanger, Norway
[7] Department of Surgery, Institute of Clinical Sciences, Sahlgrenska Academy, University of Gothenburg, Gothenburg, Sweden
[8] Estonian Genome Center, University of Tartu, Tartu, Estonia
[9] National Institute for Health and Welfare, Helsinki, Finland
[10] Division of Genetics and Epidemiology, The Institute of Cancer Research, London, United Kingdom
[11] Institute for Molecular Medicine Finland (FIMM), University of Helsinki, Helsinki, Finland
[12] Finnish Cancer Registry, Institute for Statistical and Epidemiological Cancer Research, Helsinki, Finland
[13] Faculty of Social Sciences, University of Tampere, Tampere, Finland

Cancer Genetics and Epigenetics

Key words: colorectal cancer, genetic predisposition to disease, genome-wide association study, single-nucleotide polymorphism

Abbreviations: CI: confidence interval; CRC: colorectal cancer; GWAS: genome-wide association study; LD: linkage disequilibrium; MAF: minor allele frequency; OR: odds ratio; PCA: principal component analysis; Q-Q-plot: quantile–quantile plot; SNP: single-nucleotide polymorphism

Additional Supporting Information may be found in the online version of this article.

Conflict of interest: We have no conflicts of interest to declare.

Grant sponsor: Academy of Finland (Finnish Center of Excellence Program 2012–2017); **Grant number:** 1250345; **Grant sponsor:** Cancer Society of Finland; **Grant sponsor:** Sigrid Juselius Foundation; **Grant sponsor:** Jane and Aatos Erkko Foundation; **Grant sponsor:** SYSCOL; **Grant number:** 258236; **Grant sponsor:** Nordic Information for Action eScience Center (NIASC); **Grant sponsor:** Nordic Center of Excellence financed by NordForsk; **Grant number:** 62721; **Grant sponsor:** NordForsk Colorectal Cancer Pilot Project; **Grant number:** 07 BM 11/424; **Grant sponsor:** Cancer Research UK (for Edinburgh Colon Cancer Genetics Group (CCGG)); **Grant number:** C348/A18927; **Grant sponsor:** UK Medical Research Council for Edinburgh CCGG; **Grant number:** MR/K018647/1; **Grant sponsor:** Estonian RC; **Grant numbers:** IUT20-60, PUT736; **Grant sponsor:** European Regional Development Fund; **Grant number:** 2014-2020.4.01.15-0012; **Grant sponsor:** Academy of Finland; **Grant number:** 287665; **Grant sponsor:** National Cancer Institute, USA; **Grant number:** UM1 CA167551; **Grant sponsor:** Australasian Colorectal Cancer Family Registry; **Grant numbers:** U01 CA074778, U01/U24 CA097735; **Grant sponsor:** Mayo Clinic Cooperative Family Registry for Colon Cancer Studies; **Grant number:** U01/U24 CA074800; **Grant sponsor:** Ontario Familial Colorectal Cancer Registry; **Grant number:** U01/U24 CA074783; **Grant sponsor:** Seattle Colorectal Cancer Family Registry; **Grant number:** U01/U24 CA074794; **Grant sponsor:** University of Hawaii Colorectal Cancer Family Registry; **Grant number:** U01/U24 CA074806; **Grant sponsor:** USC Consortium Colorectal Cancer Family Registry; **Grant number:** U01/U24 CA074799; **Grant sponsor:** The CCFR GWASs from National Cancer Institute, USA; **Grant numbers:** U01 CA122839, R01 CA143237, U19 CA148107

DOI: 10.1002/ijc.31076

History: Received 9 June 2017; Accepted 1 Sep 2017; Online 28 Sep 2017

Correspondence to: Lauri A. Aaltonen, Department of Medical and Clinical Genetics, Medicum, University of Helsinki, Helsinki, Finland, Tel.: +[358-2941-25595]; Fax: +[358-2941-25610], E-mail: lauri.aaltonen@helsinki.fi

[14] National Institute for Health and Welfare, Helsinki, Finland
[15] Analytic and Translational Genetics Unit, Department of Medicine, Massachusetts General Hospital, Boston, MA
[16] Program in Medical and Population Genetics, The Broad Institute of MIT and Harvard, Cambridge, MA
[17] Department of Neurology, Massachusetts General Hospital, Boston, MA
[18] Department of Surgery, Abdominal Center, Helsinki University Hospital, Helsinki, Finland
[19] Department of Pathology, Central Finland Central Hospital, Jyväskylä, Finland
[20] Department of Surgery, Jyväskylä Central Hospital, University of Eastern Finland, Jyväskylä, Finland
[21] Department of Genetics, King Faisal Specialist Hospital and Research Center, Riyadh, Saudi Arabia
[22] Wellcome Trust Centre for Human Genetics and NIHR Comprehensive Biomedical Research Centre, Oxford, United Kingdom
[23] Institute of Cancer and Genomic Sciences, University of Birmingham, Birmingham, United Kingdom
[24] Colon Cancer Genetics Group, University of Edinburgh and MRC Human Genetics Unit, Western General Hospital, Edinburgh, United Kingdom
[25] The Roslin Institute, University of Edinburgh, Easter Bush, Roslin, United Kingdom
[26] Centre for Population Health Sciences, University of Edinburgh, Edinburgh, United Kingdom
[27] Division of Cancer and Genetics, School of Medicine, Cardiff University, Cardiff, United Kingdom
[28] CRUK/MRC Oxford Institute for Radiation Oncology, University of Oxford, Oxford, United Kingdom
[29] MRC Clinical Trials Unit, , Aviation House, London, United Kingdom
[30] Oxford Cancer Centre, Department of Oncology, University of Oxford, Churchill Hospital, Oxford, United Kingdom
[31] Nuffield Department of Clinical Laboratory Sciences, John Radcliffe Hospital, University of Oxford, Oxford, United Kingdom
[32] Colorectal Oncogenomics Group, Genetic Epidemiology Laboratory, Department of Pathology, The University of Melbourne, Melbourne, VIC, Australia
[33] Centre for Epidemiology and Biostatistics, The University of Melbourne, Melbourne, VIC, Australia
[34] Cancer Prevention Program, Fred Hutchinson Cancer Research Center, Seattle, WA
[35] Lunenfeld-Tanenbaum Research Institute, Mount Sinai Hospital, Toronto, ON, Canada
[36] Department of Preventive Medicine, University of Southern California, Los Angeles, CA
[37] Department of Epidemiology and Biostatistics, Case Western Reserve University, Cleveland, OH
[38] Center for Public Health Genomics, University of Virginia, Charlottesville, VA

Genome-wide association studies have been successful in elucidating the genetic basis of colorectal cancer (CRC), but there remains unexplained variability in genetic risk. To identify new risk variants and to confirm reported associations, we conducted a genome-wide association study in 1,701 CRC cases and 14,082 cancer-free controls from the Finnish population. A total of 9,068,015 genetic variants were imputed and tested, and 30 promising variants were studied in additional 11,647 cases and 12,356 controls of European ancestry. The previously reported association between the single-nucleotide polymorphism (SNP) rs992157 (2q35) and CRC was independently replicated ($p = 2.08 \times 10^{-4}$; OR, 1.14; 95% CI, 1.06–1.23), and it was genome-wide significant in combined analysis ($p = 1.50 \times 10^{-9}$; OR, 1.12; 95% CI, 1.08–1.16). Variants at 2q35, 6p21.2, 8q23.3, 8q24.21, 10q22.3, 10q24.2, 11q13.4, 11q23.1, 14q22.2, 15q13.3, 18q21.1, 20p12.3 and 20q13.33 were associated with CRC in the Finnish population (false discovery rate < 0.1), but new risk loci were not found. These results replicate the effects of multiple loci on the risk of CRC and identify shared risk alleles between the Finnish population isolate and outbred populations.

What's new?

Genetic studies in isolated populations help characterize monogenic diseases and are being used more and more for the genetic analysis of complex diseases. Here, the authors performed a genome-wide association study with Finnish individuals afflicted with colorectal cancer. They confirm a previously reported association of a single-nucleotide polymorphism (rs992157) on chromosome 2q35, a finding independently replicated in a meta-analysis of European-ancestry individuals. Although no new risk loci were identified, the study underscores the importance of founder populations in the genetic evaluation of disease susceptibility.

Colorectal cancer (CRC) is the third most common cancer worldwide and accounts for approximately 10% of global cancer incidence and mortality (http://globocan.iarc.fr/). Numerous genetic loci have been associated with CRC in genome-wide association studies (GWASs; https://www.ebi.ac.uk/gwas/), but much of its heritability remains unexplained, which limits personalized risk assessment and biological understanding of the disease.[1,2] Discovery of new loci and replication of previously reported associations is thus important, and recent studies have continued to reveal novel CRC risk variants.[3–7] The genetic architecture of CRC varies between populations, and studies in isolated founder populations can offer valuable insights into disease susceptibility.[8]

We conducted a GWAS of CRC in the Finnish population (the FIN cohort) using a large publicly available reference panel to impute genotypes and thus increase the odds of identifying disease-associated alleles across a wide range of allele frequencies.[9] Thirty promising variants were investigated further in 11 European-ancestry studies (STHLM2, Gothenburg, HUNT, Estonia, FINRISK, COIN, UK1, Scotland1, VQ58, CCFR1 and CCFR2), adding to a total of 13,348 CRC cases and 26,438 controls.

In a recent meta-analysis of GWASs, the single-nucleotide polymorphism (SNP) rs992157 at 2q35, intronic to *PNKD* and *TMBIM1*, was found to be associated with CRC ($p = 3.15 \times 10^{-8}$; odds ratio (OR), 1.10; 95% confidence interval (CI), 1.06–1.13).[6] To replicate this finding, we genotyped and analyzed rs992157 in 4,439 CRC cases and 15,847 controls from five Northern European cohorts (STHLM2, Gothenburg, HUNT, Estonia and a subset of the FIN cohort) that had not been previously studied for the association between rs992157 and CRC.

Material and Methods

This study was conducted in accordance with the Declaration of Helsinki and approved by the Finnish National Supervisory Authority for Welfare and Health, National Institute for Health and Welfare (THL/151/5.05.00/2017) and the Ethics Committee of the Hospital District of Helsinki and Uusimaa (HUS/408/13/03/03/09). We derived 1,627 cases with colorectal adenocarcinoma from the ongoing Finnish CRC collection and genotyped normal tissues (colorectal tissue or blood) with Illumina (San Diego, CA) HumanOmni2.5–8 SNP arrays.[10,11] Illumina HumanCoreExome SNP array data for additional 91 CRC patients and 14,187 Finnish cancer-free controls were obtained from the National FINRISK Study (https://www.thl.fi/fi/web/thlfi-en/research-and-expertwork/population-studies/the-national-finrisk-study). Data on diagnosed cancers in the FINRISK study participants were collected from the Finnish Cancer Registry. PLINK v.1.90b3i (www.cog-genomics.org/plink/1.9/) was used for quality control.[12] A total of 122 samples (17 genotyped with the HumanOmni2.5–8 array and 105 genotyped with HumanCoreExome array) were excluded on the basis of close relatedness (identity-by-descent coefficient > 0.2), duplication, discordant sex information or low genotyping rate. The FIN cohort consisted of the remaining 1,701 CRC cases and 14,082 cancer-free controls. By design, the HumanOmni2.5–8 SNP array contained 2,315,673 autosomal sites, 273,074 of which overlapped with the HumanCoreExome SNP array (https://support.illumina.com/downloads.html). Exclusion criteria for SNPs were genotyping rate <95%, excess homozygosity (frequency of rare homozygotes exceeding the frequency of heterozygotes or any rare homozygous genotype with minor allele frequency (MAF) <2%), deviation from the Hardy–Weinberg equilibrium ($p < 1 \times 10^{-8}$), differential missingness between genotyping batches ($p < 1 \times 10^{-8}$), differential patterns of linkage disequilibrium (LD) in cases versus controls and LD-based strand inconsistency. After quality control, 214,705 SNPs were pre-phased with SHAPEIT v2 (r790), and genotypes were imputed with a publicly available reference panel (https://imputation.sanger.ac.uk/; http://www.haplotype-reference-consortium.org/).[9] Variants with low allele frequency (<0.4%) or low IMPUTE2 info score (<0.4) were excluded prior to association analysis. In Stage 1, disease associations were tested with a linear mixed model (BOLT-LMM-inf; https://data.broadinstitute.org/alkesgroup/BOLT-LMM/), adjusting for log-transformed age and sex.[13] A linear mixed model was used because it can control for population structure and cryptic relatedness.[14] The age covariate was defined as age at CRC diagnosis in cases and age at right censoring (end of follow-up or death) in controls. An additive genetic model was assumed. The genomic inflation factor was estimated by dividing the observed median of the BOLT-LMM-inf test statistic by the median of the chi-squared distribution with one degree of freedom. The Benjamini–Hochberg method was used to adjust for false discovery rate.

In Stage 2, the MassARRAY System by Agena Bioscience (San Diego, CA) was utilized at the Institute for Molecular Medicine Finland (FIMM) to genotype single-nucleotide variants in Nordic cohorts (STHLM2, 544 cases/541 controls; Gothenburg, 1,903 cases/258 controls; HUNT, 1,168 cases/1,147 controls; Estonia, 257 cases/259 controls; and FINRISK, 198 cases/172 controls), as well as 1,038 individuals from the FIN cohort who had also been genotyped with SNP arrays (925 with the HumanOmni2.5–8 array and 113 with the HumanCoreExome array). The STHLM2 cohort consisted of men who had been referred to prostate-specific antigen screening in Stockholm County, Sweden between 2010 and 2012; DNA samples were provided by the Karolinska Institute Biobank (http://ki.se/forskning/ki-biobank). The Gothenburg cohort was formed from CRC patients who had been operated at the Sahlgrenska University Hospital, Gothenburg, Sweden; DNA samples from cases and controls were provided by the Sahlgrenska Biobank (https://www.gothiaforum.com/sab). DNA samples from the HUNT cohort were provided by the Norwegian Nord-Trøndelag Health Study (HUNT) and Biobank (https://www.ntnu.edu/hunt). The Estonia cohort was derived from the sample collections of the Estonian Genome Center (www.geenivaramu.ee/en). The FINRISK cohort consisted of participants of the National FINRISK Study (198 CRC cases and 172 cancer-free controls) who had not been included in the FIN cohort due to unavailable SNP array data; DNA samples were provided by the THL Biobank, Finland (https://www.thl.fi/fi/web/thlfi-en/topics/information-packages/thl-biobank). When possible, cancer-free controls were matched to CRC cases on year of birth and sex. To assess imputation accuracy, squared Pearson correlation coefficients (r^2) between IMPUTE2 genotype dosage and MassARRAY genotype were calculated.

To enable standard meta-analysis, data from the FIN cohort were reanalyzed by unconditional logistic regression under an additive genetic model, adjusting for sex, log-transformed age and 10 principal components (SNPTEST v.2.5.2). In the MassARRAY-genotyped Nordic cohorts, unconditional logistic regression was applied using R v.3.3.3, provided that at least 10 minor alleles were observed. Details of the previously published GWASs (COIN, UK1, Scotland1, VQ58, CCFR1 and CCFR2) can be found in Ref. 15. Genomic control was applied by multiplying the standard errors of regression coefficients by the square root of the inflation factor of the respective study. PLINK v.1.90b3i was used for LD-based SNP pruning and principal component analysis (PCA). PCA was performed using 13,012 LD-pruned SNPs with allele frequency > 5% and IMPUTE2 info score > 0.9. R v.3.3.3 was used for meta-analysis. Estimated log ORs and standard errors were combined to obtain summary *p*-values, ORs, and 95% CIs under inverse-variance weighted random-effects and fixed-effect models (function "rma.uni" in the metafor package v.1.9-9). All reported *p*-values are two-sided. The type I error rate (α) was 0.05, corresponding to a genome-wide significance threshold of 5×10^{-8}.

Results

In Stage 1, we used a linear mixed model (BOLT-LMM-inf)[13] to test 9,068,015 single-nucleotide variants for association with CRC in the FIN cohort, which comprised 1,701 Finnish CRC cases and 14,082 population-matched, cancer-free controls. The median of the BOLT-LMM-inf test statistic was 0.512, corresponding to an inflation factor of 1.12, which was used for genomic control. A quantile–quantile (Q-Q) plot is shown in Supporting Information Figure 1, PCA plots in Supporting Information Figures 2 and 3 and a Manhattan plot in Supporting Information Figure 4. A low-frequency variant at 12q14.3 (rs73121704; MAF, 0.860%) displayed the smallest *p* value in Stage 1 ($p = 4.07 \times 10^{-9}$). Among the highest-ranking SNPs were the known CRC risk variants rs10505477 ($p = 5.29 \times 10^{-8}$), rs6589219 ($p = 4.34 \times 10^{-7}$; r^2 with rs3802842, 0.942 in 1,000 Genomes Phase 3 European populations) and rs6983267 ($p = 1.38 \times 10^{-6}$).[16–18] Thirty-eight previously published CRC risk SNPs were tested for association with CRC in the FIN cohort, and 14 of the 38 SNPs showed associations with false discovery rate < 0.1. Directions of effects were consistent with earlier publications for each of the 14 SNPs, which were located at 11q23.1 (rs3802842, $q = 1.77 \times 10^{-5}$), 8q24.21 (rs6983267, $q = 1.77 \times 10^{-5}$; rs7014346, $q = 1.77 \times 10^{-5}$), 20p12.3 (rs961253, $q = 6.92 \times 10^{-5}$), 15q13.3 (rs4779584, $q = 1.29 \times 10^{-3}$), 10q22.3 (rs704017, $q = 1.91 \times 10^{-3}$), 18q21.1 (rs4939827, $q = 7.96 \times 10^{-3}$), 2q35 (rs992157, $q = 7.96 \times 10^{-3}$), 8q23.3 (rs16892766, $q = 0.0113$), 14q22.2 (rs4444235, $q = 0.0231$), 6p21.2 (rs1321311, $q = 0.0231$), 20q13.33 (rs4925386, $q = 0.0501$), 10q24.2 (rs1035209, $q = 0.0536$) and 11q13.4 (rs3824999, $q = 0.0604$). Stage 1 results and LocusZoom plots (http://locuszoom.org/) are shown in Supporting Information Tables 1 and 2 and in Supporting Information Figures 35–102, respectively.

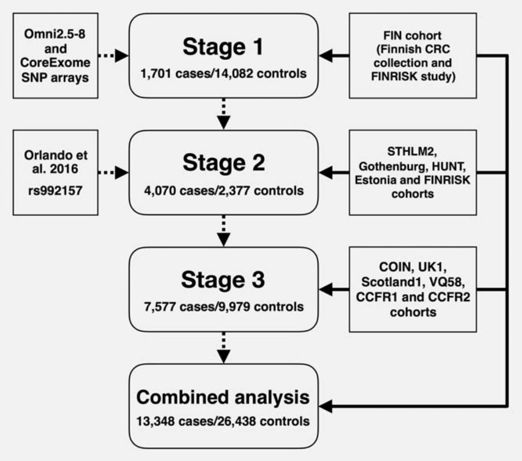
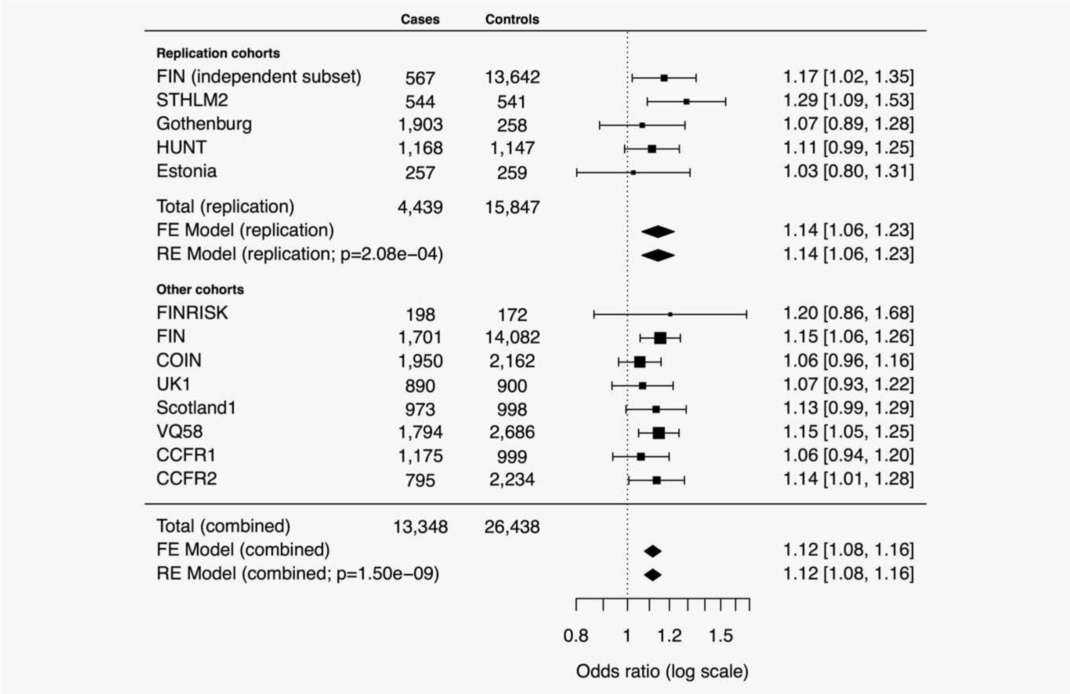

Omni2.5-8 and CoreExome SNP arrays
Stage 1
1,701 cases/14,082 controls
FIN cohort (Finnish CRC collection and FINRISK study)
Orlando et al. 2016
rs992157
Stage 2
4,070 cases/2,377 controls
STHLM2, Gothenburg, HUNT, Estonia and FINRISK cohorts
Stage 3
7,577 cases/9,979 controls
COIN, UK1, Scotland1, VQ58, CCFR1 and CCFR2 cohorts
Combined analysis
13,348 cases/26,438 controls

Figure 1. Study scheme. Sources of genetic markers are shown on the left, analytic stages in the center and sources of samples on the right.

From 20 loci that were ranked highest in Stage 1, we selected 40 variants for MassARRAY genotyping in five Nordic cohorts (STHLM2, Gothenburg, HUNT, Estonia and FINRISK; Stage 2). Two variants were selected from each locus. rs992157 (2q35) was also selected for Stage 2 because it had been recently reported as a CRC risk factor. We were unable to design genotyping assays for seven variants because of sequence context, and four variants failed genotyping. Consequently, 30 variants representing 20 loci were successfully genotyped in a total of 4,070 Nordic CRC cases and 2,377 controls. The MAF of 6:73457627G>C was low in all five Nordic cohorts, ranging from 0.000923 to 0.00954 (allele count, 2–7). To evaluate imputation accuracy, 1,038 individuals from the FIN cohort were directly genotyped with the MassARRAY platform. Squared Pearson correlation coefficients (r^2) between IMPUTE2 genotype dosage and MassARRAY genotype for the 30 variants ranged from 0.816 to 1.00 (median, 0.978).

In Stage 3, we obtained summary statistics from previously published GWASs that comprised 7,577 CRC cases and 9,979 controls of European ancestry.[15] Summary-level data were available for 27 of the 30 variants that were genotyped in Stage 2 (data for rs150509351, rs186867472 and 6:73457627G>C were missing).

To increase statistical power, datasets from Stages 1 to 3 were combined (Fig. 1), totaling 13,348 CRC cases and 26,438 controls.[19] The FIN cohort was reanalyzed by logistic regression to obtain log ORs and corresponding standard errors; the inflation factor was 1.11. The post-imputation inflation factors for the COIN, UK1, Scotland1, VQ58, CCFR1 and CCFR2 studies were 1.10, 1.03, 1.04, 1.04, 1.03 and 1.08, respectively.[15] Genomic control was applied for each of these studies. Inflation factors for the STHLM2,



	Cases	Controls	
Replication cohorts			
FIN (independent subset)	567	13,642	1.17 [1.02, 1.35]
STHLM2	544	541	1.29 [1.09, 1.53]
Gothenburg	1,903	258	1.07 [0.89, 1.28]
HUNT	1,168	1,147	1.11 [0.99, 1.25]
Estonia	257	259	1.03 [0.80, 1.31]
Total (replication)	4,439	15,847	
FE Model (replication)			1.14 [1.06, 1.23]
RE Model (replication; p=2.08e−04)			1.14 [1.06, 1.23]
Other cohorts			
FINRISK	198	172	1.20 [0.86, 1.68]
FIN	1,701	14,082	1.15 [1.06, 1.26]
COIN	1,950	2,162	1.06 [0.96, 1.16]
UK1	890	900	1.07 [0.93, 1.22]
Scotland1	973	998	1.13 [0.99, 1.29]
VQ58	1,794	2,686	1.15 [1.05, 1.25]
CCFR1	1,175	999	1.06 [0.94, 1.20]
CCFR2	795	2,234	1.14 [1.01, 1.28]
Total (combined)	13,348	26,438	
FE Model (combined)			1.12 [1.08, 1.16]
RE Model (combined; p=1.50e−09)			1.12 [1.08, 1.16]

Figure 2. Study cohorts, sample sizes and estimated odds ratios for rs992157. The vertical line corresponds to the null hypothesis (odds ratio = 1). The horizontal lines and square brackets indicate 95% confidence intervals. Areas of the boxes are proportional to the weight of the study. Diamonds represent combined estimates. FE, fixed-effect. RE, random-effects.

Gothenburg, HUNT or Estonia studies were not estimated because of the small number of genotyped markers. Fixed-effect meta-analysis was performed, but to account for possible study heterogeneity, we considered the random-effects model (Supporting Information Table 3). Under the random-effects model, rs10505477 (8q24.21), rs6983267 (8q24.21) and rs992157 (2q35) were genome-wide significant (for rs10505477, $p = 7.63 \times 10^{-14}$, $p_{\mathrm{het}} = 0.144$, $I^2 = 34.4\%$; for rs6983267, $p = 7.45 \times 10^{-13}$, $p_{\mathrm{het}} = 0.0985$, $I^2 = 37.7\%$; for rs992157, $p = 1.50 \times 10^{-9}$, $p_{\mathrm{het}} = 0.777$, $I^2 = 0\%$), and rs6589219 (11q23.1) displayed suggestive evidence of association ($p = 9.14 \times 10^{-6}$, $p_{\mathrm{het}} = 0.153$, $I^2 = 36.5\%$). Combined effect size estimates and directions of effects for these four SNPs were consistent with prior studies.[6,16–18]

Next, we studied rs992157 (2q35) in a replication dataset comprising 4,439 CRC cases and 15,847 controls (STHLM2, Gothenburg, HUNT, Estonia and a subset of the FIN cohort) who had not been previously studied for the association between rs992157 and CRC (Fig. 2). In the FIN cohort, rs992157 had been directly genotyped with SNP arrays in both cases and controls, and the other Nordic cohorts were genotyped with the MassARRAY platform. Logistic regression models were fit within each cohort. In the independent subset of the FIN cohort (567 CRC cases and 13,642 cancer-free controls), the inflation factor was 1.11, and genomic control was applied accordingly. Estimated log ORs were combined under random-effects and fixed-effect models, the results of which were highly similar without notable study heterogeneity ($p_{\mathrm{het}} = 0.462$, $I^2 = 0\%$). Applying Bonferroni correction for the 30 variants that were genotyped in the MassARRAY experiment ($\alpha = 0.05/30 \approx 0.00167$), rs992157 was significantly associated with CRC with an OR of 1.14 (95% CI, 1.06–1.23; $p = 2.08 \times 10^{-4}$). Consistent with prior results, the alternative allele (A) conferred a higher risk of CRC than the reference allele (G). For rs992157, r^2 between IMPUTE2 genotype dosage and MassARRAY genotype was 1.00 in the FIN cohort.

Discussion

The identification of CRC susceptibility alleles and quantification of their effects is biologically and clinically meaningful. The genome-wide statistical analysis of tag SNPs has highlighted new genes and regulatory mechanisms in the pathogenesis of CRC while concurrently allowing more accurate estimation of the personalized risk of colorectal neoplasms.[20,21] We conducted a GWAS of CRC in the Finnish population (Stage 1), genotyped 30 promising variants in five Nordic cohorts (Stage 2) and analyzed corresponding summary statistics from previously published GWASs (Stage 3). A total of 39,786 individuals (13,348 CRC cases and 26,438 controls) were analyzed in Stages 1–3. New genotype data generated in this study were used to analyze the recently reported effect of rs992157 (2q35) on CRC risk.

The association between rs992157 and CRC was independently replicated ($p = 2.08 \times 10^{-4}$), and its effect size was ~1.1 (OR, 1.14; 95% CI, 1.06–1.23). In the combined analysis of 13,348 CRC cases and 26,438 controls, the p value and OR for rs992157 were 1.50×10^{-9} and 1.12 (95% CI, 1.08–1.16), respectively, with no indication of study heterogeneity ($p_{\text{het}} = 0.777$, $I^2 = 0\%$). In addition to CRC, rs992157 has shown pleiotropic effects on adult human height and inflammatory bowel disease.[6,22]

In Stage 1, we found evidence supporting multiple previously published SNPs as risk factors for CRC in the Finnish population with false discovery rate < 0.1. The corresponding chromosomal regions and nearby genes were 2q35 (*PNKD* and *TMBIM1*), 6p21.2 (*TRNAI25*), 8q23.3 (*LINC00536* and *EIF3H*), 8q24.21 (*CCAT2* and *LOC101930033*), 10q22.3 (*ZMIZ1-AS1*), 10q24.2 (*NKX2–3* and *SLC25A28*), 11q13.4 (*POLD3*), 11q23.1 (*COLCA1* and *COLCA2*), 14q22.2 (*RPS3AP46* and *MIR5580*), 15q13.3 (*SCG5* and *GREM1*), 18q21.1 (*SMAD7*), 20p12.3 (*FGFR3P3* and *CASC20*) and 20q13.33 (*LAMA5*).

We did not find Finnish population-specific CRC risk variants, which may reflect limitations in replicating them in other populations, their rarity or small contributions to inherited risk. A low-frequency variant at 12q14.3 (rs73121704; MAF, 0.860%) displayed a notable association in Stage 1 ($p = 4.07 \times 10^{-9}$), but the finding was not supported by meta-analysis (random-effects $p = 0.466$, fixed-effect $p = 0.122$). Bias due to genotype imputation or population stratification remains a concern, and further data is needed.

A limitation of the study is that the number of variants selected for Stages 2 and 3 was relatively small, and disease-associated variants may have been omitted from further investigation because of low rank in the primary analysis. It is also difficult to assess whether there was residual confounding due to population stratification or different genotyping platforms. For rs992157, r^2 between IMPUTE2 genotype dosage and MassARRAY genotype was 1.00, making technical bias unlikely. Genomic control was applied for all primary GWASs to avoid type I error.

In conclusion, we replicated the association between rs992157 (2q35) and CRC in Northern European studies and found it to be genome-wide significant in a meta-analysis of 12 European-ancestry studies. SNPs at 2q35, 6p21.2, 8q23.3, 8q24.21, 10q22.3, 10q24.2, 11q13.4, 11q23.1, 14q22.2, 15q13.3, 18q21.1, 20p12.3 and 20q13.33 were associated with CRC in the Finnish population, which validates findings from previous studies and reveals shared genetic architecture of CRC between the Finnish population isolate and outbred populations.

Acknowledgements

Authors are thankful to Sini Nieminen, Sirpa Soisalo, Marjo Rajalaakso, Inga-Lill Svedberg, Iina Vuoristo, Alison Ollikainen and Heikki Metsola for their technical support. Niko Välimäki received grant No. 287665 from the Academy of Finland. Authors acknowledge Karolinska Institute Biobank, Sahlgrenska Biobank, HUNT Biobank, THL Biobank, and Estonian Genome Center for providing DNA samples from Nordic CRC cases and controls. The Nord-Trøndelag Health Study (HUNT Study) is a collaboration between HUNT Research Centre (Faculty of Medicine and Health Sciences, NTNU, Norwegian University of Science and Technology), Nord-Trøndelag County Council, Central Norway Health Authority, and Norwegian Institute of Public Health.

References

1. Graff RE, Möller S, Passarelli MN, et al. Familial risk and heritability of colorectal cancer in the Nordic Twin Study of Cancer. *Clin Gastroenterol Hepatol* 2017;15:1256–64.
2. Frampton MJE, Law P, Litchfield K, et al. Implications of polygenic risk for personalised colorectal cancer screening. *Ann Oncol* 2016;27: 429–34.
3. Zeng C, Matsuda K, Jia W-H, et al. Identification of susceptibility loci and genes for colorectal cancer risk. *Gastroenterology* 2016;150:1633–45.
4. Wang M, Gu D, Du M, et al. Common genetic variation in ETV6 is associated with colorectal cancer susceptibility. *Nat Commun* 2016;7: 11478.
5. Wang H, Schmit SL, Haiman CA, et al. Novel colon cancer susceptibility variants identified from a genome-wide association study in African Americans. *Int J Cancer* 2017;140: 2728–33.
6. Orlando G, Law PJ, Palin K, et al. Variation at 2q35 (PNKD and TMBIM1) influences colorectal cancer risk and identifies a pleiotropic effect with inflammatory bowel disease. *Hum Mol Genet* 2016;25:2349–59.
7. Schumacher FR, Schmit SL, Jiao S, et al. Genome-wide association study of colorectal cancer identifies six new susceptibility loci. *Nat Commun* 2015;6:7138.
8. Nyström-Lahti M, Kristo P, Nicolaides NC, et al. Founding mutations and Alu-mediated recombination in hereditary colon cancer. *Nat Med* 1995; 1:1203–6.
9. McCarthy S, Das S, Kretzschmar W, et al. A reference panel of 64,976 haplotypes for genotype imputation. *Nat Genet* 2016;48:1279–83.
10. Salovaara R, Loukola A, Kristo P, et al. Population-based molecular detection of hereditary nonpolyposis colorectal cancer. *J Clin Oncol* 2000;18: 2193–200.
11. Aaltonen LA, Salovaara R, Kristo P, et al. Incidence of hereditary nonpolyposis colorectal cancer and the feasibility of molecular screening for the disease. *N Engl J Med* 1998;338: 1481–7.
12. Chang CC, Chow CC, Tellier LC, et al. Second-generation PLINK: rising to the challenge of larger and richer datasets. *GigaScience* 2015;4. Available from: https://doi.org/10.1186/s13742-015-0047-8. Accessed June 9, 2017.
13. Loh P-R, Tucker G, Bulik-Sullivan BK, et al. Efficient Bayesian mixed-model analysis increases association power in large cohorts. *Nat Genet* 2015;47:284–90.
14. Pirinen M, Donnelly P, Spencer CCA. Efficient computation with a linear mixed model on large-scale data sets with applications to genetic studies. *Ann Appl Stat* 2013;7:369–90.
15. Al-Tassan NA, Whiffin N, Hosking FJ, et al. A new GWAS and meta-analysis with 1000Genomes imputation identifies novel risk variants for colorectal cancer. *Sci Rep* 2015;5: 10442.
16. Zanke BW, Greenwood CMT, Rangrej J, et al. Genome-wide association scan identifies a colorectal cancer susceptibility locus on chromosome 8q24. *Nat Genet* 2007;39:989–94.
17. Tomlinson I, Webb E, Carvajal-Carmona L, et al. A genome-wide association scan of tag SNPs identifies a susceptibility variant for colorectal cancer at 8q24.21. *Nat Genet* 2007;39: 984–8.
18. Tenesa A, Farrington SM, Prendergast JGD, et al. Genome-wide association scan identifies a colorectal cancer susceptibility locus on 11q23 and

replicates risk loci at 8q24 and 18q21. *Nat Genet* 2008;40:631–7.

19. Skol AD, Scott LJ, Abecasis GR, et al. Joint analysis is more efficient than replication-based analysis for two-stage genome-wide association studies. *Nat Genet* 2006;38:209–13.
20. Dunlop MG, Dobbins SE, Farrington SM, et al. Common variation near CDKN1A, POLD3 and SHROOM2 influences colorectal cancer risk. *Nat Genet* 2012;44:770–6.
21. Tuupanen S, Turunen M, Lehtonen R, et al. The common colorectal cancer predisposition SNP rs6983267 at chromosome 8q24 confers potential to enhanced Wnt signaling. *Nat Genet* 2009;41:885–90.
22. Wood AR, Esko T, Yang J, et al. Defining the role of common variation in the genomic and biological architecture of adult human height. *Nat Genet* 2014;46:1173–86.