Review

The missing heritability of familial colorectal cancer

Stephanie A. Schubert,* Hans Morreau, Noel F. C. C. de Miranda and Tom van Wezel*

Department of Pathology, Leiden University Medical Center, Leiden University, Albinusdreef 2, 2333 ZA Leiden, The Netherlands

*To whom correspondence should be addressed. Tel: +31 71 5266629; Fax: +31 71 5266952; Email: t.van_wezel@lumc.nl

Received 20 June 2019; Accepted 5 September 2019.

Abstract

Pinpointing heritability factors is fundamental for the prevention and early detection of cancer. Up to one-quarter of colorectal cancers (CRCs) occur in the context of familial aggregation of this disease, suggesting a strong genetic component. Currently, only less than half of the heritability of CRC can be attributed to hereditary syndromes or common risk loci. Part of the missing heritability of this disease may be explained by the inheritance of elusive high-risk variants, polygenic inheritance, somatic mosaicism, as well as shared environmental factors, among others. A great deal of the missing heritability in CRC is expected to be addressed in the coming years with the increased application of cutting-edge next-generation sequencing technologies, routine multigene panel testing and tumour-focussed germline predisposition screening approaches. On the other hand, it will be important to define the contribution of environmental factors to familial aggregation of CRC incidence. This review provides an overview of the known genetic causes of familial CRC and aims at providing clues that explain the missing heritability of this disease.

Introduction

Colorectal cancer (CRC) is a common disease; it is the third most frequent cancer in men and the second most frequent in women worldwide (1). Family history is a major risk factor for the development of CRC (2,3). Familial aggregation of this disease is observed in up to 25% of the cases (4–6). In addition, the risk for colorectal carcinomas is increased in individuals with a personal or family medical history of colorectal polyps (2,7). Thus, CRC syndromes have historically been divided into polyposis (adenomatous, hamartomatous, mixed and serrated) and non-polypsis syndromes, based largely on the number and histology of the colorectal polyps. However, their clinical presentation can be highly variable and overlapping. Luckily, molecular tumour features can aid in the diagnosis of hereditary CRC syndromes caused by underlying germline DNA repair defects, as seen in Lynch syndrome (LS) and MUTYH-associated polyposis (MAP). In both instances, tumour analysis helped identify the aetiology of these syndromes; microsatellite instability ( MSI) resulted in the identification of the DNA mismatch repair (MMR) gene MSH2 as the first causal gene for LS, and base-excision repair (BER)-deficiency pointed towards biallelic MUTYH mutations as the molecular basis of disease in a subset of colorectal adenomatous polyposis patients. Thus, the screening of somatic features in cancer tissues has already proven to be crucial in the identification of novel predisposition genes in the past and is likely to play a large role in routine onco-genetic diagnostics in the coming years with the widespread implementation of next-generation sequencing (NGS) methods in the molecular pathology. Currently, tumours can be classified more extensively on the basis of somatic alterations than previously possible, focussing on somatic mutations in genetic hotspot regions, tumour mutation burden (TMB) (i.e. the number of somatic events), and somatic mutational signatures, describing genome-wide mutation patterns generated by specific underlying mutational processes (8). NGS approaches analysing germline (i.e. leucocyte-derived) DNA have been less successful in identifying novel familial CRC genes than originally anticipated (9,10). Nevertheless, they have played a key role in the discovery of some predisposition genes for CRC and polyposis, e.g. POLE, POLD1, NTHL1 and MSH3; however, these syndromes only explain a minority of the familial cases (11).

Currently, only up to one-quarter of all the familial CRCs can be attributed to the known Mendelian-inherited syndromes, suggesting
that a substantial portion of its heritability still remains unexplained (12). In the polyposis subgroup, germline diagnostic approaches are more successful as less than 20% of the cases remain unexplained after screening the known polyposis-associated genes (13). Strikingly, no genetic cause can be identified for MMR-deficiency in 60% of the MMR-deficient CRC cases, the so-called ‘suspected Lynch syndrome’ or ‘Lynch-like syndrome’ patients (14,15). Finally, in the MMR-proficient non-polyposis familial CRC subgroup, in spite of enormous efforts in the last years, no causal genes have been confirmed. The following review is an overview of the genetic causes of familial CRCs and proposes explanations for the missing heritability of this heterogenous disease.

**Known CRC and polyposis syndromes**

Approximately 5% of the CRCs occur as a consequence of well-defined inherited syndromes (16) (Table 1). The majority of these cases are caused by CRC or polyposis syndromes, which can be either dominantly or recessively inherited, while a small fraction can be attributed to rare multi-tumour syndromes with an increased risk for CRC (12,17).

**Lynch syndrome**

LS (OMIM#120435) is the most frequent form of hereditary CRC, estimated to account for almost 3% of all CRCs in the general population (18,19). This syndrome also predisposes to endometrial, ovarian, stomach, small bowel and urinary tract cancers (20).

It is caused by heterozygous mutations in one of the DNA MMR genes, MSH1, MSH2, MSH6 and PMS2. Furthermore, deletion of EPCAM resulting in transcriptional inactivation of MSH2 and heritable MLH1 promoter hypermethylation can cause LS (21–23). Notably, biallelic germline mutations in one of the MMR genes, known as constitutional MMR-deficiency syndrome (CMMRD; OMIM#276300), predispose to a large spectrum of childhood and adolescent cancers and adenomatous polyposis (24). The MMR system broadly consists of two major components: MutS and MutL (25). MutS heterodimers (MSH2/MSH6 or MSH2/MSH3) recognise mismatches and small insertions/deletions (indels). MutL heterodimers (MLH1/PMS2, MLH1/PMS1 or MLH1/MLH3) act as endonucleases after complexing with MutS. In general, the clinical phenotype of MLH1 or MSH2 mutation carriers is more severe than PMS2 or MSH6 mutation carriers, likely due to the existence of functional redundancy within the MutS and MutL heterodimers (26,27). MMR-deficiency in LS-associated neoplasms is usually the result of a second hit in the respective MMR gene, frequently leading to loss of protein expression and a microsatellite instability-high (MSI-H) phenotype (28). Approximately 15% of the sporadic CRCs also show this hypermutator phenotype, mainly due to somatic promoter hypermethylation of the MLH1 gene (29). The MSI-H hypermutator phenotype is associated with a high mutation load (≥10 mutations/Mb) with a preferential accumulation of C:G>T:A and C:G>A:T mutations and thousands of small insertions/deletions, particularly at microsatellite sequences (9,30). This mutational process is thought to underlie multiple mutational signatures identified.

**Table 1. Colorectal cancer-associated genes**

<table>
<thead>
<tr>
<th>Genes</th>
<th>Syndrome</th>
<th>Abbr.</th>
<th>Inheritance</th>
<th>Proportion of colorectal cancer</th>
<th>Tumour genetics</th>
<th>COSMIC mutational signatures v3a</th>
</tr>
</thead>
<tbody>
<tr>
<td>MLH1</td>
<td>Lynch syndrome</td>
<td>LS</td>
<td>Dominant</td>
<td>~3%</td>
<td>MSI-H, hypermutated</td>
<td>SBS6, SBS15, SBS21, SBS26, SBS44, DBS7, DBS10, ID7, ID1; ID2</td>
</tr>
<tr>
<td>MSH2</td>
<td>(Constitutional mismatch repair-deficiency)</td>
<td>CMMRD</td>
<td>(Recessive)</td>
<td>(Unknown)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>MSH6</td>
<td>PMS2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>APC</td>
<td>(Attenuated) Familial adenomatous polyposis</td>
<td>AFAP/FAP</td>
<td>Dominant</td>
<td>~1%</td>
<td></td>
<td></td>
</tr>
<tr>
<td>MUTYH</td>
<td>MUTYH-associated polyposis</td>
<td>MAP</td>
<td>Recessive</td>
<td>&lt;1%</td>
<td></td>
<td>SBS36</td>
</tr>
<tr>
<td>POLE</td>
<td>POLE/POLD1-associated tumour syndromeb</td>
<td>PPATS</td>
<td>Dominant</td>
<td>Unknown</td>
<td>Ultra-hypermutated</td>
<td>SBS10a, SBS10b, SBS14, SBS20, DBS3</td>
</tr>
<tr>
<td>NTHL1</td>
<td>NTHL1-associated tumour syndrome</td>
<td>NATS</td>
<td>Recessive</td>
<td>Unknown</td>
<td></td>
<td>SBS30</td>
</tr>
<tr>
<td>SMAD4</td>
<td>Juvenile polyposis</td>
<td>JPS</td>
<td>Dominant</td>
<td>&lt;1%</td>
<td></td>
<td></td>
</tr>
<tr>
<td>BMPR1A</td>
<td>STK11</td>
<td>PJS</td>
<td>Dominant</td>
<td>&lt;1%</td>
<td></td>
<td></td>
</tr>
<tr>
<td>GREM1</td>
<td>Hereditary mixed polyposis syndrome</td>
<td>HMPS</td>
<td>Dominant</td>
<td>&lt;1%</td>
<td></td>
<td></td>
</tr>
<tr>
<td>MSH3</td>
<td>MSH3-associated polyposis</td>
<td></td>
<td>Recessive</td>
<td>Unknown</td>
<td>EMAST</td>
<td></td>
</tr>
<tr>
<td>RNF43</td>
<td>Serrated polyposis syndrome</td>
<td>SPS</td>
<td>Dominant</td>
<td>Unknown</td>
<td>CIMP</td>
<td></td>
</tr>
<tr>
<td>PTEN</td>
<td>PTEN hamartoma tumour syndrome</td>
<td>PHTS</td>
<td>Dominant</td>
<td>&lt;1%</td>
<td></td>
<td></td>
</tr>
<tr>
<td>TP53</td>
<td>Li-Fraumeni syndrome</td>
<td>LFS</td>
<td>Dominant</td>
<td>Unknown</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

ahttps://cancer.sanger.ac.uk/cosmic/signatures/(8).

bAlso known as polymerase proofreading-associated polyposis (PPAP).

*More than 10 000 mutations per tumour.

SBS, single base substitution signature; DBS, double base substitution signature; ID, insertion and deletion signature.
in CRCs or other tumours: single base substitution mutational signatures SBS6, SBS15, SBS21, SBS26 and SBS44, double base substitutions DBS7 and DBS10 and insertion and deletion signature ID7 and very large numbers (>10 000) of ID1 and ID2 mutations (8).

**Familial adenomatous polyposis**

The most prevalent polyposis syndrome and the second most common cause of hereditary CRC, explaining approximately 1% of the CRCs, is familial adenomatous polyposis (FAP; OMIM#175100), caused by heterozygous mutations in the APC gene (19). While FAP exhibits an autosomal dominant pattern of inheritance, approximately 25% of affected individuals have no family history and present with *de novo* mutations (31,32). FAP is a severe polyposis syndrome associated with the development of hundreds to thousands of colorectal adenomas at early age and a lifetime risk of CRC of nearly 100% (16). Attenuated FAP (AFAP) is a milder form of adenomatous polyposis with fewer adenomas (<100) at a later age, and is typically caused by APC mutations located at the proximal and distal ends of the gene or exon 9 (33,34). Both FAP and AFAP patients have an increased risk for the development of extra-colonic polypos (16). APC is a large tumour suppressor (2843 amino acids and 312 kDa) involved in the Wnt signalling pathway, which functions by negatively regulating the β-catenin oncoprotein (35). APC is somatically inactivated in about 80% of the sporadic CRCs and is believed to be one of the initiating steps of tumorigenesis (36). Therefore, the pathogenesis of FAP tumours seems to mimic their sporadic counterparts (37).

**MUTYH-associated polyposis**

Another possible diagnosis for patients presenting with 10–100 adenomas is MAP (OMIM#608456), which accounts for less than 1% of all CRCs (19). The lifetime risk of CRC development in MAP patients is ~80%, while their risk for extra-colonic cancers seems only slightly increased (16,38,39). MAP is an autosomal recessive disorder caused by biallelic mutations in the *MUTYH* gene (40,41). MUTYH is a BER protein, involved in the cell’s protection and repair mechanism against one of the most common oxidative DNA lesions, 8-oxoguanine (8-oxoG) (42). MUTYH suppresses tumorigenesis by repairing mutagenic G:C>T:A transversions, but also by inducing MUTYH-dependent cell death (40,43). Consequently, MUTYH-defective tumours show an excess of G:C>T:A transversions, which are characteristic of mutational signature SBS36 (8,44). In addition, the detection of a c.34G>T transversion at codon 12 of KRAS in colorectal tumours can point to the diagnosis of MAP (45). Furthermore, the widespread occurrence of chromosomal copy-neutral loss of heterozygosity (cnLOH) is another peculiar feature of MAP CRCs (46).

**NTHL1-associated tumour syndrome**

*NTHL1* encodes a DNA glycosylase that participates in BER and recessively inherited mutations in this gene cause a polyposis and CRC syndrome (OMIM#616415) estimated to be at least five times less frequent than MAP (47,48). More extensive clinical characterisation of biallelic mutation carriers has prompted a name change from NTHL1-associated polyposis (NAP) to the broader NTHL1-associated tumour syndrome (NATS), as these patients are also at increased risk of developing brain, breast and some other cancer types (10,47). Similar to MAP tumours, the BER-deficiency in NATS tumours results in a unique, clearly distinct mutational signature composed of G:C>T:A transitions at non-CpG sites, designated as mutational signature SBS30 (8,49).

**POLE/POLD1-associated tumour syndrome**

A small fraction of familial CRC patients are carriers of a heterozygous germline mutation in the exonuclease domain of the DNA polymerase subunits POLE (OMIM#150833) or POLD1 (OMIM#612591); however, their exact prevalence remains unknown (50,51). Originally identified in adenomatous polyposis patients, carriers have also been found amongst non-polyposis CRC patients with MMR-proficient or MMR-deficient tumours (52,53). Their tumour spectrum has recently been extended to include additional neoplasms, such as endometrial, ovarian, brain and pancreatic cancer (54–56). The original name coined for this syndrome is ‘polymerase proofreading-associated polyposis (PPAP)’; however, we believe ‘POLE/POLD1-associated tumour syndrome’ would be a more appropriate name due to the recently expanded tumour spectrum and the absence of colonic polyposis in some patients. The genes POLE and POLD1 code for the catalytic subunit of DNA polymerase ε (epsilon) and δ (delta), involved in DNA replication of the leading and lagging strands, respectively (57). Tumours with pathogenic POLE or POLD1 mutations, either germline or somatic, show an ultra-hypermutator phenotype with the number of somatic mutations exceeding the 100 mutations/Mb and suggestive of a deficiency of the proofreading capacity of these polymerases (9,58). POLE-mutant tumours especially show an excess of C:G→A:T and C:G→T:A mutations, characteristic of mutational signatures SBS10a and SBS10b, and enrichment of DBS3 in some tumours (8,30). Thus far, no clear mutational signatures have been associated with POLD1 exonuclease domain mutations. Interestingly, a combined deficiency in MMR function and POLE or POLD1 proofreading is relatively common and results in the generation of unique mutational signatures SBS14 and SBS20, respectively (8).

**Rare polyposis syndromes**

A fraction of the familial CRC risk can be explained by rare autosomal dominant hamartomatous polyposis syndromes, occurring with approximately one-tenth of the frequency of adenomatous polyposis syndromes (59). Two of the most common are juvenile polyposis (JPS; OMIM#174900; SMAD4 and BMPR1A) and Peutz-Jeghers syndrome (PJS; OMIM#175200; STK11) (60–63). In addition, patients with hereditary mixed polyposis syndrome (HMPS; OMIM#601228; GREM1) present with polyps of different histological type, including hamartomatous polyps (59,64). These syndromes have distinct clinical phenotypes, histological features, frequencies and location of polyps, organ-specific manifestations and predispositions for the development of other malignancies (16,60).

In addition, in very rare cases, serrated polyposis syndrome (SPS) can be attributed to heterozygous (likely) pathogenic mutations in the RNF43 gene (OMIM#17108; 7 families), encoding an E3 ubiquitin ligase that acts as a Wnt inhibitor (65,66). A common feature of the serrated pathway is the occurrence of CpG island methylator phenotype (CIMP), which was also observed in more than 75% of the RNF43 syndrome-associated colonic lesions (66). Another very rare genetic cause of adenomatous polyposis was reported in one study, where biallelic mutations in the MMR gene *MSH3* were identified in two families with unexplained adenomatous polyposis, resulting in MSH3-associated polyposis syndrome (OMIM#617100) (67). These tumours showed loss of MSH3 expression and a high degree of microsatellite instability in tetranucleotide repeats (EMAST). MSH3-deficiency has previously been shown to induce this mutator phenotype (68,69). The exact prevalence and penetrance of

---

---
heterozygous mutations in RNF43 and biallelic mutations in MSH3 still remain unknown and warrant further research.

**Multi-tumour syndromes**

Currently, approximately three-quarters of familial CRC cases cannot be explained by germline mutations in known CRC-associated genes (12). Both classical testing strategies and multigene panel tests in CRC cases have uncovered pathogenic germline variants in genes associated with inherited syndromes not primarily linked with an increased risk for CRC.

**PTEN hamartoma tumour syndrome**

Heterozygous germline mutations in the tumour suppressor gene PTEN can cause PTEN hamartoma tumour syndrome (PHTS; OMIM#158350), a collection of rare syndromes with overlapping clinical presentations (70). The phenotypic spectrum of PHTS includes Cowden syndrome (CS), Bannayan-Riley-Ruvalcaba syndrome (BRRS), Proteus syndrome (PS) and Proteus-like syndrome (PSL) (70). CS and BRRS are considered variable phenotypic presentations of the same hamartomatous polyposis syndromes (71). CS is the best-described phenotype of PHTS and is classically characterised by the development of multiple hamartomas and characteristic dermatological manifestations, but also with an increased risk of breast, endometrial, thyroid and CRC (60,72). BRRS patients are generally affected by gastrointestinal hamartomatous polyps, lipomas, haemangiomas and macrocephaly, but a high cancer risk has been less well-documented (60,71).

**Li-Fraumeni syndrome**

Another multi-tumour syndrome with an increased risk for CRC is Li-Fraumeni syndrome (LFS; OMIM#151623), caused by heterozygous mutations in the TP53 gene. This is a very severe hereditary cancer syndrome with a lifetime cancer risk of 70–100% (73). Patients are predisposed to various and multiple early-onset cancers, mainly bone and soft tissue sarcomas, breast cancer and brain cancer (74). Between 4–14% of the Li-Fraumeni syndrome patients are diagnosed with CRC, mostly before 50 years of age (74–76). Furthermore, recent germline screening efforts in large cohorts of early-onset CRC patients frequently (0.1–1%) detected TP53 mutations in these patients (19,77,78).

**Other hereditary tumour syndromes**

Recently, additional genes causing hereditary tumour syndromes not classically associated with CRC have also been implicated in CRC predisposition (11). For example, hereditary breast cancer genes (OMIM#11448; e.g. BRCA1, BRCA2, CHEK2, ATM and PALB2) might confer a low-to-moderate risk for CRC (19,79). Rare CHEK2 variants have been associated with a 2-fold increased risk of CRC (80). CHEK2 is considered a low-to-moderate risk gene in multi-organ cancer susceptibility (OMIM#604373), including breast, thyroid, prostate, colon and kidney tumours (81). The CHEK2 mutation c.1100delC is primarily associated with an increased risk for breast cancer, but it seems especially frequent in families with both hereditary breast and CRC combined (82,83). Furthermore, CRC risk associations in BRCA1 (OMIM#60437) and BRCA2 (OMIM#612555) carriers have yielded contradictory results with risk increases up to 5-fold (84,85). However, a recent meta-analysis found a 1.5-fold higher risk of CRC in BRCA1 mutation carriers, but none in BRCA2 mutation carriers (86). Other genes linked to CRC risk include breast and pancreatic cancer genes ATM (OMIM#607585) and PALB2 (OMIM#610355), and melanoma and pancreatic cancer gene CDKN2A (OMIM#60719) (19,79,87). It is likely that with the routine implementation of multigene panel tests in the clinic more phenotypic overlaps will be observed between traditionally distinct cancer predisposition syndromes, resulting in the recognition of additional CRC risk genes in the coming years.

**Common CRC risk loci**

The ‘common-disease common-variant’ model attributes some genetic susceptibility of common diseases to the inheritance of many risk alleles common in the population, which individually confer low-risk increases, but combined they can predispose an individual to a substantial risk (88). For CRC predisposition, approximately 100 common low-risk loci have been identified in genome-wide association studies (GWASs) to date, and validated risk loci include chromosome 8q24.21, 10q22.3, 12p13.13 and 15q13.3, associated with a 1.04- to 1.23-fold increased risk (89–91). Some of the genes located in these regions have already been linked to CRC aetiology, such as components of the transforming growth factor beta (TGF-β) signalling pathway, including GREM1, SMAD3, SMAD7, SMAD9, BMP2, BMP4 and RHPN2 (92). The TGF-β signalling pathway is implicated in tumorigenesis due to its biological role in cell proliferation, differentiation, migration and apoptosis and the presence of a transcriptional signature characteristic of TGF-β activation in a large group of colorectal tumours with a poor prognosis (consensus molecular classification (CMS) 4 subgroup) (93,94). In addition, rare pathogenic germline mutations in the TGF-β genes SMAD4, BMPRIA and GREM1 are responsible for hereditary polyposis syndromes. The combined risk of inheriting multiple low-risk alleles can be estimated using risk prediction models, as so-called polygenic risk scores (PRS). More research has been conducted on PRS in individual clinical management for breast cancer; however, recent studies have hinted at a possible implementation in the clinic for familial CRC patients (89,90,95). Enrichment of low-risk loci has also been observed in familial CRC research (96). The combined contribution of low-risk variants to the familial risk is estimated to be up to 15% (90,91,97,98). These studies also indicate that more research is needed to identify additional uncommon susceptibility alleles through expanded sample sizes and increased ethnic diversity. Furthermore, research has mainly focussed on the identification of low-risk single nucleotide variants (SNVs), while the characterisation of common gross chromosomal structural variations and their influence on cancer risks remains largely unexplored (99,100).

**Candidate variants for CRC**

Comprehensive genetic screens were intensively applied in the last decade to discover novel genetic causes of CRCs and were able to identify some promising candidates; however, it is likely that additional culprits will be identified in the future.

**Monogenic inheritance**

Up to half of the missing heritability of CRC might be caused by still elusive high-risk monogenic variants (101). NGS approaches have specifically boosted the identification of candidate genes in smaller families or families suspected of lower penetrance mutations, in which genetic linkage analysis failed to identify causal loci. To date, multiple candidate genes for hereditary CRC and polyposis
have been proposed, including genes encoding proteins involved in DNA repair, DNA damage response and Wnt signalling (11). The most promising candidate genes, identified in two or more studies, are GALTN12 (11 families), BUB1 (6 families), RPS20 (3 families) and AXIN2 (3 families) (102–112). Additional candidate genes include FAN1 (3 families), LRP6 (2 family), SETD6 (1 family) and MCM9 (1 family), among others, identified in single whole-exome sequencing studies (113–116). Furthermore, a combination of homozygosity mapping and sequencing strategies identified MIA3 as a novel candidate gene (117). The problem with the identification of novel cancer predisposition genes using family-based genetic screening approaches is the reproducibility in validation cohorts. A systematic review of eight candidate genes was only able to validate variants in one gene (RPS20), but not in the other genes (FANCM, FAN1, TP53, BUB1, BUB3, LRP6 and PTPN12) (106). The study by Chubb et al., where 16% of familial CRC cases could be attributed to high-risk mutations, found that the majority of the variants were identified in the most prevalent CRC-associated genes such as the MMR genes and APC (118). These studies refute the existence of novel high-risk susceptibility genes that are recurrently mutated in unrelated populations and they point towards variants which will only explain the hereditary predisposition for individuals within few or single (extended) families.

Polygenic inheritance

The co-inheritance of multiple rare low-to-moderate risk alleles likely cause a substantial proportion of hereditary cancers where high-risk variants cannot explain familial segregation of disease (88). However, these variants are difficult to identify with the current research approaches focussing on either common low-risk variants or rare high-risk variants in the familial context. To date, few rare moderate-risk variants have been identified in CRC-associated genes (16,80). The increased risk for monoallelic carriers of pathogenic MUTYH variants is still under debate, although available data suggests an at most 3-fold increased risk of CRC (119). The APC c.I1307K variant, carried by ~6% of Ashkenazi Jews, has been shown to increase both the risk of colorectal adenomas and cancer by approximately 2-fold (120). Furthermore, a digenic inheritance model, were the genotype at two loci explains the phenotype, was proposed once in an early-onset CRC patient carrying heterozygous germline variants in two genes involved in the repair of oxidised DNA lesions (MUTYH and OGG1) (121,122). Additionally, the co-occurrence of MSH6 and MUTYH within CRC patients points towards a synergistic effect of monoallelic variants in both genes (123,124).

Missed genetics

The presently implemented diagnostic screening approaches are missing and misclassifying some pathogenic genetic alterations. Broader diagnostic screening approaches and technological improvements of NGS platforms are likely to aid in the identification of these unexplored causal variants; however, to what extent these variants can explain the missing heritability of CRC is currently unknown.

Somatic mosaicism

Part of the heritability of CRC can already be explained by somatic mosaicism, the phenomenon in which variants are only present in a fraction of an organism’s cells. The first report of somatic mosaicism in CRC patients was two decades ago in apparent de novo FAP patients (125). Current estimates show that 20% of de novo FAP cases have somatic APC mosaicism (126,127). Depending on the embryonic stage at which a somatic mutation occurs, i.e. whether it affects gonadal tissues, the mutation can be transmitted to the offspring (128,129). Screening strategies which include DNA derived from multiple tissues are more sensitive and specific than leuco- cyte DNA analysis alone in the identification of mosaic APC mutations (128). Little research has been performed on the prevalence and penetrance of mosaic mutations in other CRC-associated genes. However, more mosaic patients are likely to be identified in the future, as tumour-focussed predisposition screening approaches will probably be more widely applied in clinical diagnostics. Thus far, two families have been described with a mosaic MMR mutation, one with MSH2 and the other with MLH1 mosaicism (130,131).

Epigenetic inheritance

Constitutional epimutations are epigenetic defects that cause disruption of gene expression. These epimutations are present in all somatic tissues and can occur due to aberrant promoter hypermethylation in the absence or presence of underlying genetic changes, known as primary and secondary epimutations, respectively. The existence of transmission of primary epimutations, i.e. transgenerational epigenetic inheritance, is still under debate in humans, due to intergenerational DNA methylation erasure (132). The inheritance of an epimutation was first described in a LS family with constitutional MSH2 promoter hypermethylation (133). In this family and multiple additional patients EPCAM deletions were subsequently identified, which cause the atypical epigenetic state of MSH2, classifying them as secondary epimutations (21,22). In addition, constitutional epimutations of MLH1 explain 2–3% of the suspected LS patients with MLH1-deficient tumours (134). Most of these mutations are considered primary epimutations (seemingly occurring independently of any DNA sequence changes), and cause a severe LS phenotype in carriers, but do not appear to be transmitted to the offspring (135,136). There are three reports suggesting non-Mendelian transmission of a primary epimutation, indicating that the inheritance of a primary epimutation might be possible in extremely rare cases (135–137). On the other hand, in most families suspected of a heritable constitutional MLH1 epimutation an underlying genetic defect has been identified, and therefore undiscovered genetic changes may still be present in the cases suspected of transgenerational epigenetic inheritance (135,138). Next to MLH1 other tumour suppressor genes are frequently somatically silenced by promoter hypermethylation in tumours, and constitutional epimutations of these genes, such as BRCA1 and RB1, might also predispose to cancer (139,140). It is conceivable that CRC predisposition in some patients may be caused by primary or secondary constitutional epimutations, and this warrants further exploration.

Technological limitations

For several decades, the gold standard for diagnosing CRC syndromes was Sanger sequencing of the genes matching the clinical phenotype of the patient under investigation. In recent years, short-read NGS technologies (e.g. Illumina and Ion Torrent) have been implemented in routine germline diagnostic testing, improving the detection of pathogenic mutations by screening several genes simultaneously (141). These approaches focus on exonic genomic regions and it is therefore conceivable that uncharacterised pathogenic variants are missed in regions outside the coding sequence, such as intergenic region, promoter sequences, intronic regions or
untranslated regions (UTRs). Although, whole gene capture of the MMR genes in suspected LS patients (negative for MMR mutations after diagnostic screening) was not able to identify any meaningful variants, as none of the numerous intronic variants was functionally tested (53). The strengths of the short-read sequencing technologies lie in the identification of single nucleotide variants (SNV) and small insertion/deletions (indels), while structural variant identification (large duplications, large deletions, translocations and inversions) remains challenging but not impossible. This can be illustrated by the study from Franch-Expósito et al. where a 400 kb duplication was identified by whole-exome sequencing analysis (142). Also, paralogous sequence variant discrimination can be complicated with the current sequencing analyses, e.g. identification of pathogenic variants in PMS2, one of the genes associated with LS, is hampered by multiple pseudogenes (143). Furthermore, mutation identification is troublesome in high GC-content regions and repetitive sequences, i.e. short tandem repeat (STR) and variable number tandem repeat (VNTR). In the future, the routine implementation of novel sequencing technologies such as long-read sequencing [e.g. PacBio SMRT sequencing and Oxford Nanopore Technology (ONT)] will likely overcome some of these limitations (144). In addition, an alternative for DNA diagnostic strategies is transcriptome sequencing [RNA sequencing (RNA-seq), which screens for expressed germline variants, splice-site variants, chimeric gene fusions and, possibly, (epi)mutations affecting gene expression (145)]. Nevertheless, a combination of these approaches will most likely be necessary to elucidate the remaining heritability of CRC.

Variants of unknown significance

A possible drawback of sequencing larger regions of the genome is the identification of more variants of unknown significance (VUS). The International Agency for Research on Cancer (IARC) classification system divides genetic variants into five classes: class 5 (pathogenic), class 4 (likely pathogenic), class 3 (uncertain), class 2 (likely not pathogenic) and class 1 (not pathogenic). For instance, the large majority of the missense variants identified in the MMR genes during LS screening, lack sufficient corresponding clinical data and are therefore classified as VUS (class 3) (146). To address this issue high-throughput functional assays have been developed for the diagnostic assessment of variants identified in suspected LS patients, examples are RNA splicing assay, cell-free MMR assays (CIMBRA), protein heterodimer formation assays (Yeast 2 hybrid assay) and protein localisation experiments (147–150). More recently, the functional assessment of splice-site variants was performed on RNA isolated from formalin-fixed paraffin-embedded tissue (151). Due to the complexity of the data and the large clinical consequences for patients, variant classification will probably need to be performed by a panel of experts based on an integrated assessment of the different functional assays, in silico prediction tools, clinical and epidemiological data. The next challenge will be the development and implementation of similar functional assessment protocols used for LS for VUS identified in other CRC-associated genes.

Non-genetic risk factors

In developed countries, there is a steady decline in the incidence of CRC, largely attributed to the widespread introduction of screening programmes, but since the past two to three decades the incidence of early-onset disease (diagnosed <50 years) is rapidly increasing, representing a challenge for the healthcare system as this group of patients consists mostly of sporadic cases with late-stage disease (152,153). As CRC is 3-fold more frequently diagnosed in more developed countries compared to less developed countries, this points towards the influence of dietary patterns, obesity and lifestyle factors on cancer risk (1). The World Cancer Research Fund/American Institute for Cancer Research reports strong evidence for an increased risk of CRC for the consumption of red and processed meat, alcohol intake, obesity and height, while physical activity, consumption of whole-grains, fibres, dairy products and calcium supplements shows strong evidence for a risk decrease (154). Greater adherence to these recommendations, including non-smoking, has been shown to reduce CRC risk (155,156). Mounting evidence suggests that the effect of dietary patterns on CRC risk is, at least partially, mediated by the gut microbiota composition, i.e. microbial diversity and colonisation of specific bacterial strains (157,158). The correlation of CRC with the gut microbiome is a relatively recent discovery, and the effect of the microbiome, but also the use of probiotics and antibiotics, on CRC risk is still intensely researched. Additional, non-environmental, risk factors for CRC include age, male gender and chronic intestinal inflammation, among others (1,159). Patients with a personal history of inflammatory bowel disease (IBD), a term used to describe ulcerative colitis and Crohn disease, have an increased risk for CRC, especially those with controlled disease with severe, longstanding inflammation (160–162). Genetic variations have been identified that are associated with IBD, and these loci may also underlie some of the familial aggregation of CRC (163).

An explanation for part of the missing heritability of CRC might lie in the overestimation of the actual heritability of disease. Shared non-genetic factors amongst close relatives can result in the belief that multiple cancers within the same family are the result of an underlying genetic defect, while, in fact, the cancer predisposition may be explained by exposure to shared environmental factors (159). However, shared environmental factors in adulthood alone do not primarily attribute to familial clustering, as CRC risk estimates between spouses is less than 1.1 (164). Furthermore, in families with exactly two first-degree relatives affected by a given disease there is only a moderate probability of this being due to the segregation of genetic variants (165). In addition, the familial risk might be overestimated in relatives of suspected LS patients. In approximately half of these patients, the MSI-H phenotype can be explained by two pathogenic somatic mutations, highlighting an additional mechanism resulting into MMR-deficiency next to germline mutations, constitutional epimutations and somatic hypermethylation of the MLH1 promoter (166,167). Risk estimates in these families should, therefore, be based on family history, as an underlying pathogenic germline variant outside the known LS-associated genes cannot be excluded, such as germline mutations in POLE, POLD1 or MUTYH (52,168).

Conclusions

Only 25% of all familial CRCs can be explained by rare pathogenic mutations in one of the CRC-associated genes (12,169). The combined effect of common low-risk loci is currently estimated to be up to 15% of the familial risk, but this is likely to increase once less common variants and structural aberrations are included in large-scale analyses (90). Taken together, more than half of the CRC heritability still remains unexplained. Still undiscovered monogenic cancer predisposition genes may explain up to half of the missing heritability of CRC. Cancer risk in additional familial cases might result from the polygenic inheritance of multiple uncharacterised moderate-risk variants. Furthermore, the current diagnostic screening approaches are likely to be missing some of the pathogenic
genetic alterations, such as somatic mosaicism and gross chromosomal variations. Variants of unknown significance (VUS) probably explain some of the remaining CRC heritability. In other cases, the seeming heritability might not be genetic, but may in fact, result from shared environmental risk factors. Future research will likely focus on the integrated analysis of an individual’s environmental exposures with his or her genetic background to study gene-environment interactions.

Hopefully, in the coming years broader diagnostic screening approaches and technological improvements can partially elucidated the missing heritability of CRC. Increasing evidence shows phenotypic overlap between CRC and non-CRC syndromes, supporting multigene panel testing over strict germline testing of only syndrome-specific genes, which has the potential to miss a sizable fraction of the hereditary cancer patients (19,170,171). Lastly, the implementation of mutational signature and TMB analyses in molecular pathology is impending, as its importance as a biomarker for response to novel immunotherapies is becoming more apparent (172,173). A substantial proportion of the genes and candidate genes associated with familial CRC encode proteins involved in DNA repair mechanisms, resulting in distinct somatic mutation patterns in these tumours. Therefore, tumour-focussed germline predisposition screening approaches will likely play an increasing role in oncogenetic diagnostics by aiding in the identification of known and potentially novel hereditary cancer syndromes.

**Funding**

This work was supported by Dutch Digestive Foundation (grant number MLDS FP13-11 to T.v.W.) and the Fight Colorectal Cancer-Michael’s Mission-AACR Fellowship in Young Onset, Late-Stage Colorectal Cancer Research 2015 (grant number 15-40-1645-DEMI to N.F.C.C.d.M.).

**Acknowledgements**

This article is based upon work from COST Action CA17118, supported by COST (European Cooperation in Science and Technology). Conflict of interest statement: None declared.

**References**


Heritability of familial colorectal cancer

69. Carethers, J. M., Koi, M. and Tseng-Rogenski, S. S. (2015) EMAST is a form of microsatellite instability that is initiated by inflammation and modulates colorectal cancer progression. Genes, 6, 185–205.


