Systematic meta-analyses, field synopsis and global assessment of the evidence of genetic association studies in colorectal cancer

Zahra Montazeri,1 Xue Li,2 Christine Nyiraneza, 1 Xiangyu Ma,3 Maria Timofeeva,4 Victoria Svinti,4 Xiangrui Meng,2 Yazhou He,2 Yacong Bo,5 Samuel Morgan,1 Sergio Castellvi-Bel,6 Clara Ruiz-Ponte,7 Ceres Fernández-Rozadilla,7 Angel Carracedo,7 Antoni Castells,6 Timothy Bishop,8 Daniel Buchanan,9,10,11 Mark A Jenkins,9 Temitope O Koku,12 Annika Lindblom,13,14 Fränzle J van Duijnhooven,15 Anna Wu,16 Susan M Farrington,4 Malcolm G Dunlop,4 Harry Campbell,2 Evropi Theodoratou, Evropi Theodoratou

ABSTRACT

Objective To provide an understanding of the role of common genetic variations in colorectal cancer (CRC) risk, we report an updated field synopsis and comprehensive assessment of evidence to catalogue all genetic markers for CRC (CRCgene2).

Design We included 869 publications after parallel literature review and extracted data for 1063 polymorphisms in 303 different genes. Meta-analyses were performed for 308 single nucleotide polymorphisms (SNPs) in 158 different genes with at least three independent studies available for analysis. Scottish, Canadian and Spanish data from genome-wide association studies (GWAS) were incorporated for the meta-analyses of 132 SNPs. To assess and classify the credibility of the associations, we applied the Venice criteria and Bayesian False-Discovery Probability (BFDP). Genetic associations classified as ‘positive’ and ‘less-credible positive’ were further validated in three large GWAS consortia conducted in populations of European origin.

Results We initially identified 18 independent variants at 16 loci that were classified as ‘positive’ polymorphisms for their highly credible associations with CRC risk and 59 variants at 49 loci that were classified as ‘less-credible positive’ SNPs; 72.2% of the ‘positive’ SNPs were successfully replicated in three large GWAS consortia and the ones that were not replicated were downgraded to ‘less-credible positive’ reducing the ‘positive’ variants to 14 at 11 loci. For the remaining 231 variants, which were previously reported, our meta-analyses found no evidence to support their associations with CRC risk.

Conclusion The CRCgene2 database provides an updated list of genetic variants related to CRC risk by using harmonised methods to assess their credibility.

INTRODUCTION

Colorectal cancer (CRC) is one of the most commonly diagnosed malignancies and the leading cause of cancer deaths in the world, with 1.65 million new cases and almost 835,000 deaths in 2015. The global burden of CRC is expected to increase by 60%, with more than 2.2 million new cases and almost 1.1 million deaths occurring annually by 2030. The distribution of CRC global burden varies widely, with more than two-thirds...
of all cases and about 60% of all deaths occurring in countries with a high or very high human development index, including Australia and New Zealand, Europe and North America, while incidence and mortality rates in Africa and South-Central Asia are relatively low. These geographic differences appear to be attributable to the differences in both environmental exposures and the background of genetically determined susceptibility.

It is estimated that 15%–25% of CRC risk variance is attributed to inherited genetic factors, and the first-degree relatives of CRC patients have two to four times higher risk of developing CRC. The inherited genetic risk of CRC can be partly accounted for by a combination of rare high-penetrance mutations and large numbers of common genetic variants each conferring small risk. Although a number of highly penetrant mutations (eg, DNA mismatch repair genes, APC, SMAD4, LKB1/STK11, MUTYH) have been identified to influence CRC susceptibility with large effects, overall they account for only 2%–5% of incident CRC cases in the general population because these mutations are very rare. Candidate gene association studies have investigated the role of a large number of common genetic variants in CRC risk, but only a small number of them have been successfully replicated in subsequent studies.

In 2012 and 2014, we reported two independently conducted series of meta-analyses to systematically evaluate associations between CRC and common variants using data from candidate gene studies and genome-wide association studies (GWASs) and identified a number of promising genetic risk variants for CRC risk. Our first field synopsis (published in 2012) reported 16 variants in 13 independent loci (MUTYH, MTHFR, SMAD7, 8q24, 8q23.3, 11q23.1, 14q22.2, 1q41, 20p12.3, 20q13.33, 3q26.2, 16q22.1, 19q13.1), and the second field synopsis (published in 2014) identified 8 additional variants in 5 independent loci (APC, CHEK2, DNMT3B, MLH1, MUTYH) that were strongly associated with CRC. These two synopses used slightly different methodologies, in that the 2012 field synopsis only included variants reported in four or more studies in meta-analyses, whereas the 2014 included variants with three or more studies, and there were some differences in the criteria applied for the evidence appraisal.

In this study, we aimed to perform an updated field synopsis for CRC risk by including the most recently published genetic association studies, by following established guidelines and using harmonised methods for evidence appraisal. We systematically captured all published genetic association data on CRC for meta-analyses and subsequently incorporated data from GWAS consortia for interrogation. This study provides an up-to-date and publicly available database for CRC genetics (CRCgene2) and presents these data within a defined statistical and causal inference framework to aid interpretation of the results. We aimed to provide new insights into the fundamental biological mechanisms involved in colorectal carcinogenesis.

**METHDOSES**

**Literature search and data collection**

To identify genetic association studies of CRC risk, we searched the Medline database via the Ovid gateway and the search terms comprising medical subject headings (MeSH) and keywords relating to colorectal neoplasms, the MeSH heading ‘genetic predisposition to disease’, and the keywords ‘geneS’ and ‘associateS’ were applied to terms in the entire article. The latest literature search was performed on 21 November 2018. We screened the eligibility of retrieved publications in a three-step parallel review of title, abstract and full text by following the predefined inclusion and exclusion criteria. Thus, each eligible study evaluated the association between a polymorphic genetic variant (with minor allele frequency ≥0.01 in the reference panel of the 1000 genomes) and a sporadic CRC. Studies investigating only premalignant conditions such as adenomas, polyps or dysplastic tissue were excluded. Studies investigating hereditary CRC syndromes, such as familial adenomatous polyposis, hereditary non-polyposis CRC, juvenile polyposis syndrome and Gardner’s syndrome; solely focusing on the progression or histological phenotype of CRC; or studies in animals; were excluded. Case-control, cohort and GWASs were included, while family-based studies were excluded. All included studies were published in English in a peer-reviewed journal; studies only reported as conference abstracts were excluded. Data from the eligible studies were abstracted into two standardised tables, including the key variables with regard to the study identifiers and context, study design and limitations, genotype information and outcome effects.

A list of genetic variants that were investigated in meta-analyses was summarised and data from three GWAS consortia (Scotland, Canada and Spain) were incorporated for meta-analyses, when the genotype data are available for the listed variants. In brief, the Study of CRC in Scotland (SOCCS) is a population-based case-control GWAS that includes 3417 cases and 3500 controls. The Assessment of Risk for Colorectal Tumors In Canada (ARCTIC) is a case-control GWAS database that includes 1231 cases and 1240 controls. The population-based cohort study in Spain comprised two phases (EPICOLON I and EPICOLON II) adding up to 2000 cases and 2000 controls. Restricted candidate gene genotyping data were available from both phases and GWAS data were only accessible from 881 cases and 667 controls from phase 2. More details about these GWAS datasets are present in online supplementary text.

**Statistical analysis**

Meta-analysis was performed for genetic variants with data available from at least three independent studies. Summary crude ORs and 95% CI for allelic, recessive and dominant genetic models were calculated by applying either the fixed-effect model (Mantel-Haenszel method) or the random-effects model (DerSimonian-Laird method) in case of the existence of substantial heterogeneity. The Q statistic (with a threshold of p value <0.05) and I² metric were calculated to quantify between-study heterogeneity. Funnel plot analysis with an Egger test was conducted to test for small study effects. We also estimated the statistical power of each meta-analysis based on the significance level of α=0.05, the effect sizes and the allele frequencies of genetic variants (an integral component of the Bayesian False-Discovery Probability (BFDP) analysis). All statistical analysis was conducted by using R software (R x64 3.1.0).

**Credibility of the identified genetic associations**

We first applied the BFDP and the Venice criteria to assess the credibility of any observed genetic associations with p<0.05 in at least one genetic model. We then validated these associations in three additional GWAS consortia: Genetics and Epidemiology of Colorectal Cancer Consortium (GECCO), Colorectal Transdisciplinary Study (CORECT, https://research. fhcrc.org/peters/en/corect-study.html) and Colon Colorectal Cancer Family Registry (CFR). With meta-analysis of these three GWAS datasets, we validated the observed genetic associations using data from 58 131 CRC cases and 67 347 controls.
Systematic review of genetic association studies on CRC risk

A meta-analyses for genetic variants with data from at least 3 studies and incorporation of data from 3 GWASs (SOCCS, ARCTIC and EPICOLON)

The Venice criteria and BFDP were applied to assess the credibility of observed genetic associations

Validation of genetic variants classified as "positive" and "less-credible positive" in 3 additional GWASs (GECCO, CORRECT and CCFR).

**Figure 1** Diagram of the study design. ARCTIC, Assessment of Risk for Colorectal Tumors In Canada; BFDP, Bayesian False-Discovery Probability; CFR, Colorectal Cancer Family Registry; CORRECT, Colorectal Transdisciplinary Study; CRC, colorectal cancer; EPICOLON, Gastrointestinal Oncology Group of the Spanish Gastroenterological Association; GECCO, Genetics and Epidemiology of Colorectal Cancer Consortium; GWASs, genome-wide association studies; SOCCS, Study of CRC in Scotland.

---

(online supplementary text), and the statistical power of validation was estimated accordingly.

The BFDP assesses the noteworthiness of an observed association. The BFDP was selected rather than the false-positive report probability (FPRP) because it uses more information, defines the noteworthiness threshold explicitly in terms of the costs of false discovery and non-discovery, and does not suffer from the inferential limitations identified for the FPRP.\(^\text{12}\) We calculated BFDP values at two levels of prior probability: a medium/low prior level (0.05 to 10\(^{-4}\)), close to what would be expected for a candidate gene, and a very low prior level (10\(^{-4}\) to 10\(^{-6}\)), close to what would be expected for a random SNP. A noteworthy threshold was defined as 0.2 based on the assumption that the cost of false discovery would be four times higher than that of false non-discovery.\(^\text{19}\)

According to the Venice criteria, the credibility of associations is assessed for three aspects: the amount of evidence, the extent of replication and the protection from bias.\(^\text{12, 15}\) We used statistical power to assess the volume of evidence and a grade of A, B and C was assigned, respectively, when statistical power was greater than 80%, 50%–79% or less than 50%. The extent of replication was assessed by the measurement of heterogeneity (I\(^2\) criterion), and a grade of A, B and C was assigned, respectively, when I\(^2\) was less than 25%, 25%–49% or greater than 50%.\(^\text{12}\)

For protection from bias, a complete assessment is difficult; instead, we considered the following aspects: (1) the phenotype definition was addressed by our inclusion criteria—namely that cases would have newly incident CRC; (2) genotyping error rates are generally low; (3) the criterion of replication across studies in part addresses potential concerns about variation in genotyping quality between studies; and (4) the magnitude of effect of population stratification appears to be small in general.\(^\text{23}\)

Genetic associations were then classified into four categories based on the following criteria. Associations were classified as ‘positive’ if they: (1) were statistically significant at a p value level of 0.05 in at least two of the genetic models, (2) had a BFDP less than 0.20 at least at the p value level of 0.05, (3) had a statistical power greater than 80%, (4) had an I\(^2\) less than 50%. A class of ‘less-credible positive’, with a less-stringent threshold, was assigned to the associations (1) that were statistically significant at a p value threshold of 0.05 in at least one of the genetic models, but (2) their BFDP was greater than 0.20 or their statistical power was between 50% and 79% or had an I\(^2\) greater than 50%. Associations with p value large than 0.05 were further classified as ‘null’ or ‘negative’ by assessing if there are more than 5000 cases. After credibility assessment, genetic variants classified as ‘positive’ and ‘less-credible positive’ were sent to the CORECT coordinating centre to validate their associations with CRC risk using additive, dominant and recessive models. At this stage, ‘positive’ associations that failed to be validated (at p<0.05) were downgraded to ‘less-credible positive’. A schematic diagram is shown in **Figure 1** to demonstrate datasets included in each phase of the analysis.

---

**RESULTS**

**Literature search and data collection**

A total of 20,900 citations were identified from literature search. Of these, 6,770 (32.4%) papers were published after the search period of the most recent field synopsis (31 December 2012).\(^\text{11}\)

After eligibility screening, we finally included and extracted data from 869 publications (**Figure 2**), reporting the association of CRC risk with 1063 polymorphisms in 303 different genes, of which 308 polymorphisms were reported in at least three independent studies.

**Meta-analyses**

Meta-analyses were conducted for 308 polymorphisms in 158 different loci with data available in three or more candidate or GWA studies (online supplementary table 1). On average, these meta-analyses were based on 6149 CRC cases (median; IQR=2301–7334) and 7337 controls (median; IQR=2809–8855) originating from 8 (median; IQR=4–9) case-control studies. Data from the Scottish, Canadian and/or Spanish GWAS were incorporated in the meta-analyses for 132 SNPs. Summary crude ORs and 95%CI for the allelic, dominant and recessive models are presented in **Table 1**. Of the meta-analyses for 308 polymorphisms (tagged at different 158 loci), a total of 77 SNPs (25.6%) (tagged in 61 different loci) were identified to have a nominally statistically significant association (p value <0.05) with CRC risk in at least one of the three genetic models and were eligible for credibility assessment using the BFDP (online supplementary tables 2–4) and the Venice criteria\(^\text{12, 15}\) and for validation in the three (GECCO, CORRECT and CCFR) GWAS consortia (online supplementary tables 5–7).

Credibility assessment indicated 18 variants (5.8% of the meta-analysed SNPs) tagging 16 loci (rs36053993 and rs34612342 in MUTYH, rs2066847 in NOD2, rs12953717 and rs44641148 in SMAD7, rs1569686 in DNMT3B, rs2736100 in TERT, rs9858822 in PPAR-gamma, rs1862748 in CDH1, rs7259371 in RHPN2, rs355527 in BMP2, rs1800469 in TGFβ1, rs10505477 in 8q24, rs16892766 in 8q23.3, rs3802842 in 11q23.1, rs961253 in 20p12.3, rs10795668 in 10p14, rs4951291 in PPAR-gamma, rs1862748 in CDH1, and rs355527 in BMP2).
1q32.1) had the most credible associations with CRC risk and are therefore referred to as ‘positive’ SNPs (table 1, online supplementary tables 2–4). These findings are based on accrued data on 1224 to 43,652 cases and on 1381 to 60,883 controls, with a median of 17,100 cases per meta-analysis. The linkage disequilibrium (LD) between these ‘positive’ polymorphisms was checked pairwise using the Ensembl LD calculator with reference to the 1000 genomes: phase 3 CEU population and we found two pairs of SNPs (rs355527 and rs961253, rs12953717 and rs4464148) with $r^2 > 0.20$ (online supplementary tables 8). The other 59 variants (19.2% of the meta-analysed SNPs) in 49 loci with p value <0.05 were classified as ‘less-credible’ positive’ SNPs, given high heterogeneity, low statistical power or a high possibility of being false positive (BFDP >0.2) for their association with CRC risk (table 2, online supplementary tables 2–4). The summary findings for ‘less-credible’ positive’ SNPs were based on accrued data on 246 to 51,730 CRC cases and on 399 to 53,589 controls, with a median of at least 4,287 CRC cases per meta-analysis.

Polymorphisms classified as either ‘positive’ or ‘less-credible positive’ SNPs were sent for validation in synthesised data from the GECCO, CORECT and CCFR consortia. Validation was only able to be performed for 68 out of 77 polymorphisms, as one ‘positive’ variant (rs34612342 in MUTYH [Y179C] gene) was dropped due to its low imputation quality in these GWASs; 5 ‘less-credible positive’ SNPs (in GSTT1, GSTM1, TP73, CHEK2 and CYP2E1) had no rs numbers; and 3 polymorphisms (in TP53 and KRAS) were not available in these GWASs; therefore, they were not able to be validated. Of the 17 ‘positive’ SNPs sent for validation, 7 polymorphisms (41.2%) in 6 different loci (rs12953717 and rs4464148 in SMAD7, rs355527 in BMP2, rs10505477 in 8q24, rs961253 in 20p12.3, rs16892766 in 8q23.3, rs3802842 in 11q23.1) reached a genome-wide statistical significance ($p\leq5\times10^{-8}$) in at least one meta-analysis model and 6 polymorphisms (35.3%) in 6 different loci (rs7259371 in RHPN2, rs2736100 in TERT, rs10795668 in 10p14, rs36053993 in MUTYH, rs1862748 in CDH1 and rs1800469 in TGFBI) reached a nominally statistical significance ($p<0.05$) in at least one meta-analysis model. However, the remaining 4 polymorphisms (23.5%) in 4 different loci (rs2066847 in NOD2, rs1569686 in DNMT3B, rs9858822 in PPAR-gamma and rs4951291 in 1q32.1) failed in the GWAS validation of all genetic models (online supplementary tables 5–7) and they were therefore downgraded to ‘less-credible positive’. Of the 51 ‘less-credible positive’ SNPs sent for validation, 2 polymorphisms (3.9%) in 2 different loci (rs6983267 in 8q24 and rs1801155 in APC) reached a genome-wide statistical significance ($p\leq5\times10^{-8}$) in at least one meta-analysis model and 11 polymorphisms (21.6%) in 11 different loci (rs1136410 in PARP1, rs11568820 in VDR, rs1342387 in ADIPOR1, rs719725 in TPDS2L3, rs20417 in PTGS2/COX2, rs2665802 in GH1, rs4803455 in TGFBI, rs7849 in SCD, rs8752 in HPGD, rs36053993 in MUTYH and rs928554 in ESR2) reached a nominally statistical significance ($p<0.05$) in at least one meta-analysis model, whereas the remaining 38 polymorphisms (74.5%) failed in the GWAS validation of all genetic models (online supplementary tables 5–7). Overall, 26 (33.8%) out of the 77 nominally significant polymorphisms tested via meta-analysis were successfully validated in these GWASs ($p<0.05$).

Funnel plots were produced for all ‘credible’ (online supplementary figures 1–14) and ‘less-credible’ SNPs (online supplementary figures 15–77) for their summary estimates in allelic analysis model. Small study effects were reported for 2 (14.3%) of the 14 ‘positive’ polymorphisms (rs36053993 in MUTYH and rs4464148 in SMAD7) and for 10 (15.9%) of 63 ‘less-credible’ polymorphisms (rs1229984 in ADH1B, rs25487 in XRCC1, rs2240308 in AXIN2, rs1799750 in MMP1, rs1048943 in CYP1A1, rs373572 in RAD18, rs1800566 in NQO1, rs2066844 in NOD2, rs2665802 in GH1, rs712 in KRAS). Therefore, their reported association with CRC risk should be interpreted with caution.

Figure 2 The distribution of included studies published from 1991 to 2018.
Table 1 Summary of ‘credible positive associations’ from meta-analyses and credibility assessment

<table>
<thead>
<tr>
<th>Gene</th>
<th>Variant</th>
<th>Cases vs controls (number of samples)</th>
<th>Allelic model</th>
<th>Heterogeneity</th>
<th>Power</th>
<th>Recessive model: var/var vs wt/wt and wt/var</th>
<th>Dominant model: wt/var and var/var vs wt/wt</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>Effective size</td>
<td></td>
<td></td>
<td>Effect size</td>
<td>Effect size</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Heterogeneity</td>
<td>Heterogeneity</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>P value</td>
<td>P value</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>I² (95% CI)</td>
<td>I² (95% CI)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>P value</td>
<td>P value</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Classification</td>
<td>Classification</td>
</tr>
<tr>
<td>MUTYH</td>
<td>rs36053993</td>
<td>28 310 vs 20 935 (19*)</td>
<td>1.42 (1.22 to 1.66)</td>
<td>4.89E-06</td>
<td>0 (0 to 3.3)</td>
<td>0.703</td>
<td>0.99</td>
</tr>
<tr>
<td>MUTYH</td>
<td>rs34612342</td>
<td>28 180 vs 20 923 (19*)</td>
<td>1.89 (1.40 to 2.42)</td>
<td>5.72E-07</td>
<td>0 (0 to 6.2)</td>
<td>0.282</td>
<td>1.00</td>
</tr>
<tr>
<td>SMA7</td>
<td>rs29572717</td>
<td>28 030 vs 20 519 (13*)</td>
<td>1.11 (0.70 to 1.64)</td>
<td>6.24E-07</td>
<td>49 (0 to 9.3)</td>
<td>0.242</td>
<td>1.00</td>
</tr>
<tr>
<td>SMA7</td>
<td>rs464548</td>
<td>17 722 vs 17 356 (9*)</td>
<td>1.12 (0.98 to 1.29)</td>
<td>2.66E-09</td>
<td>12 (0 to 9.2)</td>
<td>0.337</td>
<td>1.00</td>
</tr>
<tr>
<td>Nq24</td>
<td>rs1505477</td>
<td>18 562 vs 20 132 (14)</td>
<td>1.14 (1.10 to 1.19)</td>
<td>2.04E-11</td>
<td>17</td>
<td>1.89 (1.47 to 2.42)</td>
<td>5.72E-07</td>
</tr>
<tr>
<td>Zfp212_J</td>
<td>rs641253</td>
<td>22 971 vs 25 270 (14*)</td>
<td>1.11 (1.06 to 1.16)</td>
<td>1.58E-05</td>
<td>60 (28 to 91)</td>
<td>0.002</td>
<td>1.00</td>
</tr>
<tr>
<td>Nq24</td>
<td>rs6812766</td>
<td>8351 vs 8878 (6*)</td>
<td>1.24 (1.15 to 1.34)</td>
<td>1.67E-08</td>
<td>4 (0 to 9.2)</td>
<td>0.393</td>
<td>0.99</td>
</tr>
<tr>
<td>Igg41</td>
<td>rs1759668</td>
<td>16 760 vs 18 312 (16)</td>
<td>0.88 (0.83 to 0.94)</td>
<td>4.35E-05</td>
<td>67 (34 to 96)</td>
<td>6.34E-05</td>
<td>1.00</td>
</tr>
<tr>
<td>Itg3.1</td>
<td>rs1802942</td>
<td>22 320 vs 22 945 (20*)</td>
<td>1.15 (1.10 to 1.20)</td>
<td>2.12E-12</td>
<td>43 (0 to 7.5)</td>
<td>0.242</td>
<td>1.00</td>
</tr>
<tr>
<td>BMP2</td>
<td>rs553527</td>
<td>13 539 vs 14 173 (38)</td>
<td>1.12 (0.98 to 1.27)</td>
<td>1.65E-10</td>
<td>0 (0 to 6)</td>
<td>0.377</td>
<td>0.99</td>
</tr>
<tr>
<td>CDH1</td>
<td>rs862748</td>
<td>17 436 vs 18 416 (11)</td>
<td>0.91 (0.88 to 0.94)</td>
<td>1.98E-08</td>
<td>0 (0 to 63)</td>
<td>0.703</td>
<td>0.98</td>
</tr>
<tr>
<td>RMM2</td>
<td>rs25971711</td>
<td>15 778 vs 16 700 (9*)</td>
<td>0.87 (0.82 to 0.93)</td>
<td>1.65E-05</td>
<td>52 (0 to 9.0)</td>
<td>0.034</td>
<td>1.00</td>
</tr>
<tr>
<td>TERT</td>
<td>rs2748000</td>
<td>16 176 vs 18 125 (8)</td>
<td>1.07 (1.04 to 1.10)</td>
<td>2.95E-05</td>
<td>0 (0 to 82)</td>
<td>0.526</td>
<td>0.88</td>
</tr>
<tr>
<td>TGFBR1</td>
<td>rs1809469</td>
<td>40 271 vs 60 244 (10)</td>
<td>0.89 (0.80 to 0.99)</td>
<td>0.036</td>
<td>53 (20 to 92)</td>
<td>0.023</td>
<td>1.00</td>
</tr>
</tbody>
</table>

*Includes GWAS data from SOCCS, GWAS, genome-wide association studies; SOCCS, Study of CRC in Scotland.
<table>
<thead>
<tr>
<th>Gene</th>
<th>Variant</th>
<th>Cases vs controls (number of samples)</th>
<th>Effect size</th>
<th>Recessive model: var/var vs wt/wt and wt/var</th>
<th>Dominant model: wt/wt vs var/var vs wt</th>
<th>Power</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Allelic model</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MTHFR</td>
<td>rs1801131</td>
<td>43 (62) vs 60 (83)</td>
<td>0.96 (0.93 to 0.99)</td>
<td>3.61E-09</td>
<td>7.28E-05</td>
<td>9 (3.0 to 9.8)</td>
</tr>
<tr>
<td>miR</td>
<td>rs851129</td>
<td>132 (26) vs 161 (44)</td>
<td>1.19 (1.06 to 1.30)</td>
<td>0.033</td>
<td>1.70E-03</td>
<td>1.00 (9.2 to 9.9)</td>
</tr>
<tr>
<td>ADHIE</td>
<td>rs2299349</td>
<td>44 (69) vs 69 (95)</td>
<td>1.05 (0.94 to 1.18)</td>
<td>0.217</td>
<td>1.27E-02</td>
<td>0.000 (9.2 to 9.9)</td>
</tr>
<tr>
<td>HCC</td>
<td>XPC</td>
<td>rs229471</td>
<td>12 (11) vs 18 (16)</td>
<td>0.017</td>
<td>1.24E-02</td>
<td>0.002 (9.2 to 9.9)</td>
</tr>
<tr>
<td>miR</td>
<td>rs895819</td>
<td>112 (14) vs 164 (1)</td>
<td>1.19 (1.06 to 1.30)</td>
<td>0.033</td>
<td>1.70E-03</td>
<td>1.00 (9.2 to 9.9)</td>
</tr>
<tr>
<td>miR</td>
<td>rs1229984</td>
<td>44 (69) vs 69 (95)</td>
<td>1.05 (0.94 to 1.18)</td>
<td>0.217</td>
<td>1.27E-02</td>
<td>0.002 (9.2 to 9.9)</td>
</tr>
<tr>
<td>miR</td>
<td>rs179953</td>
<td>12 (11) vs 18 (16)</td>
<td>1.19 (1.06 to 1.30)</td>
<td>0.033</td>
<td>1.70E-03</td>
<td>1.00 (9.2 to 9.9)</td>
</tr>
<tr>
<td>miR</td>
<td>rs1800947</td>
<td>12 (11) vs 18 (16)</td>
<td>1.19 (1.06 to 1.30)</td>
<td>0.033</td>
<td>1.70E-03</td>
<td>1.00 (9.2 to 9.9)</td>
</tr>
<tr>
<td>miR</td>
<td>rs25487</td>
<td>112 (14) vs 164 (1)</td>
<td>1.19 (1.06 to 1.30)</td>
<td>0.033</td>
<td>1.70E-03</td>
<td>1.00 (9.2 to 9.9)</td>
</tr>
<tr>
<td>miR</td>
<td>rs1801131</td>
<td>132 (26) vs 161 (44)</td>
<td>1.19 (1.06 to 1.30)</td>
<td>0.033</td>
<td>1.70E-03</td>
<td>1.00 (9.2 to 9.9)</td>
</tr>
<tr>
<td>miR</td>
<td>rs1800947</td>
<td>12 (11) vs 18 (16)</td>
<td>1.19 (1.06 to 1.30)</td>
<td>0.033</td>
<td>1.70E-03</td>
<td>1.00 (9.2 to 9.9)</td>
</tr>
<tr>
<td>Gene</td>
<td>Variant</td>
<td>Cases vs. controls (number of samples)</td>
<td>Effect size</td>
<td>Heterogeneity</td>
<td>Dominant model: wT/wT and var/var vs wT/wT</td>
<td>Recessive model: var/var vs wT/wT and wt/wt</td>
</tr>
<tr>
<td>---------</td>
<td>---------</td>
<td>----------------------------------------</td>
<td>------------------------------</td>
<td>---------------</td>
<td>-------------------------------------------</td>
<td>-------------------------------------------</td>
</tr>
</tbody>
</table>
Table 2  Continued

<table>
<thead>
<tr>
<th>Gene</th>
<th>Variant</th>
<th>Cases vs controls (number of samples)</th>
<th>N</th>
<th>Alleric model</th>
<th></th>
<th>Recessive model: var/var vs wt/wt and wt/var</th>
<th>Dominant model: wt/var and var/var vs wt/wt</th>
<th>Power</th>
<th>Effect size OR (95% CI)</th>
<th>P value</th>
<th>Heterogeneity I 2 (95% CI)</th>
<th>P value</th>
<th>Power</th>
<th>Effect size OR (95% CI)</th>
<th>P value</th>
<th>Heterogeneity I 2 (95% CI)</th>
<th>P value</th>
<th>Power</th>
<th>Effect size OR (95% CI)</th>
<th>P value</th>
<th>Heterogeneity I 2 (95% CI)</th>
<th>P value</th>
</tr>
</thead>
<tbody>
<tr>
<td>ADIPOQ</td>
<td>rs2241766</td>
<td>1517 vs 1909 (5)</td>
<td>5</td>
<td>1.17 (1.01 to 1.37)</td>
<td>0.033</td>
<td>21 (0 to 97)</td>
<td>0.231</td>
<td>0.50</td>
<td>5</td>
<td>1.04 (0.74 to 1.45)</td>
<td>0.830</td>
<td>32 (0 to 96)</td>
<td>0.325</td>
<td>0.06</td>
<td>5</td>
<td>1.26 (1.09 to 1.47)</td>
<td>0.002</td>
<td>8 (0 to 96)</td>
<td>0.364</td>
<td>0.92</td>
<td>Less credible</td>
<td></td>
</tr>
<tr>
<td>CHEK2</td>
<td>rs1787996</td>
<td>1667 vs 3370 (5)</td>
<td>3</td>
<td>1.47 (1.12 to 1.92)</td>
<td>0.005</td>
<td>0 (0 to 96)</td>
<td>0.499</td>
<td>0.53</td>
<td>na</td>
<td>na</td>
<td>na</td>
<td>na</td>
<td>na</td>
<td>3</td>
<td>1.48 (1.13 to 1.95)</td>
<td>0.005</td>
<td>0 (0 to 96)</td>
<td>0.496</td>
<td>0.81</td>
<td>Less credible</td>
<td></td>
<td></td>
</tr>
<tr>
<td>CHEK2</td>
<td>1100delC</td>
<td>2417 vs 3615 (6)</td>
<td>4</td>
<td>1.66 (1.01 to 2.73)</td>
<td>0.048</td>
<td>0 (0 to 83)</td>
<td>0.835</td>
<td>0.60</td>
<td>na</td>
<td>na</td>
<td>na</td>
<td>na</td>
<td>na</td>
<td>4</td>
<td>1.66 (1.01 to 2.75)</td>
<td>0.047</td>
<td>0 (0 to 83)</td>
<td>0.835</td>
<td>0.86</td>
<td>Less credible</td>
<td></td>
<td></td>
</tr>
<tr>
<td>TP53</td>
<td>rs42522</td>
<td>10591 vs 12763 (30)</td>
<td>30</td>
<td>1.11 (0.99 to 1.24)</td>
<td>0.077</td>
<td>85 (94 to 95)</td>
<td>4.4E-26</td>
<td>0.73</td>
<td>30</td>
<td>1.21 (1.02 to 1.44)</td>
<td>0.026</td>
<td>70 (55 to 88)</td>
<td>2.7E-09</td>
<td>0.98</td>
<td>30</td>
<td>1.11 (0.96 to 1.28)</td>
<td>0.163</td>
<td>82 (84 to 05)</td>
<td>7.5E-20</td>
<td>0.64</td>
<td>Less credible</td>
<td></td>
</tr>
<tr>
<td>TP53</td>
<td>rs787162</td>
<td>1812 vs 2319 (6)</td>
<td>6</td>
<td>0.98 (0.76 to 1.36)</td>
<td>0.859</td>
<td>70 (23 to 96)</td>
<td>0.005</td>
<td>0.06</td>
<td>6</td>
<td>1.20 (0.81 to 1.72)</td>
<td>0.282</td>
<td>0 (0 to 8)</td>
<td>0.638</td>
<td>0.22</td>
<td>6</td>
<td>0.95 (0.71 to 1.27)</td>
<td>0.734</td>
<td>71 (25 to 96)</td>
<td>0.004</td>
<td>0.40</td>
<td>Less credible</td>
<td></td>
</tr>
<tr>
<td>KRAS</td>
<td>rs712</td>
<td>1992 vs 2194 (3)</td>
<td>3</td>
<td>1.22 (0.96 to 1.52)</td>
<td>0.270</td>
<td>90 (55 to 100)</td>
<td>6.8E-05</td>
<td>0.19</td>
<td>3</td>
<td>1.72 (1.13 to 0.73)</td>
<td>0.215</td>
<td>87 (44 to 100)</td>
<td>3.7E-04</td>
<td>0.16</td>
<td>3</td>
<td>1.17 (0.84 to 1.66)</td>
<td>0.350</td>
<td>83 (22 to 98)</td>
<td>0.003</td>
<td>0.28</td>
<td>Less credible</td>
<td></td>
</tr>
<tr>
<td>ARKTS1</td>
<td>rs3831855</td>
<td>2281 vs 3196 (5)</td>
<td>5</td>
<td>1.09 (0.71 to 1.18)</td>
<td>0.026</td>
<td>0 (0 to 76)</td>
<td>0.083</td>
<td>0.35</td>
<td>5</td>
<td>1.08 (0.90 to 1.29)</td>
<td>0.492</td>
<td>43 (0 to 93)</td>
<td>0.133</td>
<td>0.21</td>
<td>5</td>
<td>1.18 (0.95 to 1.46)</td>
<td>0.313</td>
<td>62 (90 to 96)</td>
<td>0.012</td>
<td>0.77</td>
<td>Less credible</td>
<td></td>
</tr>
<tr>
<td>TgL2.1</td>
<td>rs4951291</td>
<td>15 835 vs 16 724 (34)</td>
<td>9</td>
<td>0.91 (0.84 to 0.99)</td>
<td>0.021</td>
<td>68 (22 to 88)</td>
<td>0.002</td>
<td>0.85</td>
<td>9</td>
<td>0.74 (0.62 to 0.87)</td>
<td>2.7E-04</td>
<td>0 (0 to 8)</td>
<td>0.616</td>
<td>0.95</td>
<td>9</td>
<td>0.92 (0.83 to 1.01)</td>
<td>0.077</td>
<td>72 (30 to 90)</td>
<td>4.3E-04</td>
<td>0.91</td>
<td>Less credible</td>
<td></td>
</tr>
<tr>
<td>PPAR-gamma</td>
<td>rs8588222</td>
<td>2512 vs 2630 (6)</td>
<td>4</td>
<td>1.40 (1.19 to 1.65)</td>
<td>6.4E-05</td>
<td>0 (0 to 90)</td>
<td>0.697</td>
<td>0.96</td>
<td>2</td>
<td>1.69 (1.19 to 2.39)</td>
<td>0.003</td>
<td>0 (0 to 98)</td>
<td>0.801</td>
<td>0.54</td>
<td>4</td>
<td>1.47 (1.19 to 1.83)</td>
<td>4.5E-04</td>
<td>0 (0 to 98)</td>
<td>0.618</td>
<td>1.00</td>
<td>Less credible</td>
<td></td>
</tr>
<tr>
<td>DNMT3B</td>
<td>rs1569686</td>
<td>1224 vs 1381 (5)</td>
<td>4</td>
<td>0.57 (0.47 to 0.66)</td>
<td>1.8E-09</td>
<td>0 (0 to 90)</td>
<td>0.992</td>
<td>1.00</td>
<td>4</td>
<td>0.93 (0.54 to 1.58)</td>
<td>0.756</td>
<td>0 (0 to 98)</td>
<td>0.414</td>
<td>0.06</td>
<td>5</td>
<td>0.46 (0.25 to 0.83)</td>
<td>2.4E-09</td>
<td>30 (90 to 9)</td>
<td>0.223</td>
<td>1.00</td>
<td>Less credible</td>
<td></td>
</tr>
<tr>
<td>NOD2</td>
<td>rs2068147</td>
<td>4573 vs 3733 (10)</td>
<td>10</td>
<td>1.39 (1.16 to 1.66)</td>
<td>3.2E-04</td>
<td>0 (0 to 71)</td>
<td>0.547</td>
<td>0.78</td>
<td>10</td>
<td>2.36 (0.69 to 8.13)</td>
<td>0.173</td>
<td>0 (0 to 76)</td>
<td>0.720</td>
<td>0.31</td>
<td>10</td>
<td>1.38 (1.14 to 1.66)</td>
<td>0.001</td>
<td>0 (0 to 76)</td>
<td>0.488</td>
<td>0.96</td>
<td>Less credible</td>
<td></td>
</tr>
</tbody>
</table>

*Includes GWAS data from Ontario.  
†Tomlinson 2008 was based on 10 samples.  
‡Includes GWAS data from SOCCS.  
GWAS = genome-wide association studies; SOCCS = Study of CRC in Scotland.
DISCUSSION

This systematic, comprehensive field synopsis of genetic association studies on CRC updates the two previous field synopses24 to11 with application of harmonised methods for evidence appraisal and further validation of the identified genetic associations in three GWAS consortia. Specifically, we extracted and collated data for 1063 polymorphisms in 303 different genes from 869 publications and performed up-to-date meta-analyses for 308 variants in 158 different genes that had data from at least three independent studies available for analysis. After credibility assessment and validation, we identified a total of 12 genetic loci credibly associated with CRC risk, of which 6 loci (MUTYH, SMAD7, 8q24, 8q23.3, 11q23.1, 20p12.3) were also classified as credibly associated with CRC risk in the previous field synopses and the other 6 loci (TGFb1, TERT, CDH1, RHPN2, BMP2 and 10p14) are novel findings as they have not been assessed or reported as credible risk loci in the previous field synopses.

We note that a synopsis was undertaken of literature on genetic associations with CRC published in the period 2012–2017.24 Our analysis differs from that study, in that it includes and updates data from our previous field synopses,10 11 includes data from three GWAS studies in the meta-analysis and further validates the findings in three GWA consortia.

Similar to our previous field synopses,10 11 the present study reported two SNPs at 8q24 locus (rs6983267 and rs10505477) with strong evidence supporting significant associations with CRC risk and with these associations further replicated in three GWAS consortia. In biopsies of the rectum, sigmoid colon and cecum mucosa, proliferation has been reported to be higher among homozygotes for the risk alleles of the rs6983267 and rs10505477 variants compared with those with other genotypes in the general population.25 The rs6983267 variant has been assessed as having a highly credible association with CRC risk because the number of cases was less than 5000 for which the null results could be due to limited statistical power, while another 83 SNPs with p>0.05 and the number of cases>5000 were classified as null variants with adequate statistical power (online supplementary table 9).

Highly credible associations were also reported for four variants tagging four different genes that are involved in inflammation or immune response. First, a positive association with the rs3802842 variant in the 11q23.1 region was identified for CRC risk and this association was replicated in the data from the GWAS consortia and in the recent East Asian analysis.46 Fine-mapping and bioinformatics analysis performed within the GECCO-CCFR consortium supported this variant with strong functional evidence.46, 47

The remaining 231 polymorphisms (75.0%) assessed in meta-analyses were reported to have no statistically significant association (p>0.05) with CRC risk in all three genetic models (online supplementary table 9), based on accrued data on 192 to 219 292 cases and 251 to 24 054 controls, with median of at least 4017 cases per meta-analysis. Of them, 148 polymorphisms were classified as having negative associations with CRC risk because the number of cases was less than 5000 for which the null results could be due to limited statistical power, while another 83 SNPs with p>0.05 and the number of cases>5000 were classified as null variants with adequate statistical power (online supplementary table 9).

Two genetic polymorphisms (rs12953717 and rs4464148) tagging in SMAD7 were identified to be associated with CRC with highly credible evidence. Associations with rs12953717 and rs4464148 were successfully validated in the three GWAS consortia. In fine-mapping and bioinformatics analysis performed within the GECCO-CCFR consortium, rs4464148 was appraised as having less strong functional evidence than the highly correlated rs9932005 variant, which is located within 5 kb away.28 The SMAD7 protein is an inhibitor for the TGF-ß signalling pathway.42 43 There were highly credible associations with the rs1862748 (tagging CDH1), rs355257 (BMP2), rs961253 (BMP2) and rs7259371 (RHPN2) variants, which are TGF-ß related, and replicated in the data from the GWAS consortia. In the recent East Asian analysis, the association with the rs961253 variant was replicated, as well as additional variants of SMAD7 (rs7229639, rs4939827, CDH1 (rs9922918), BMP2 (rs4813802) and RHPN2 (rs10411210)).46 In the analysis from the GECCO, CORECT and CCFR consortia, conditioned on the rs4813802 and rs189583 variants of BMP2 as well as each other, novel associations with the BMP2 variants rs28488 and rs994308 were detected.46

Additionally, two variants (rs34612342 and rs36053993) tagging the MUTYH gene were highly credibility associated with increased CRC risk, of which rs36053993 was validated in the three GWAS consortia data, while rs34612342 was not tested due to its poor imputation quality. The MUTYH gene is known to be involved in the dysfunction of base-excision repair, which is the major pathway for repairing oxidative damage. This biological pathway in which the MUTYH gene is involved contributes to the development of multiple colorectal adenomas and carcinomas (MUTYH-associated polyposis (MAP) syndrome).46 In our analysis, the p values were nominally significant for all models for the rs34612342 variant and for the allelic and dominant models for the rs36053993 variant. For the former variant, the magnitude of effect was greatest for the recessive model, although with wide CIs, which is intriguing as MAP, in which highly penetrant mutations are implicated, and is inherited in an autosomal recessive manner.45

Highly credible associations were also reported for four variants tagging four different genes that are involved in inflammation or immune response. First, a positive association with the rs3802842 variant in the 11q23.1 region was identified for CRC risk and this association was replicated in the data from the GWAS consortia and in the recent East Asian analysis.46 Fine-mapping and bioinformatics analysis performed within the GECCO-CCFR consortia support this variant with strong functional evidence.46 Fine mapping identified two genes COLCA1 and COLCA2 arranged on opposite strands and sharing a regulatory region containing rs3802842.46 It is reported that carrying the risk allele of rs3802842 is associated with the expression levels of COLCA1 and COLCA2, which is further correlated with lymphocyte infiltration of colonic lamina propria. Further, in an
expression quantitative trait locus analysis in colorectal tissue, there were signals for COLCA1 and COLCA2. The polymorphism rs10795668 in 10p14 locus tagging GATA3 gene was reported with a highly credible association with CRC, and this association was further replicated in the GWAS consortia and in the recent East Asian analysis. However, in fine-mapping and bioinformatics analysis, this variant presents with weak functional evidence. Another less-credible positive association with the rs9858822 variant of PPARγ was identified; however, this association was not replicated in the GWAS consortia. Bioinformatics analysis showed that PPARγ and its ligands have been found to block proinflammatory genes in colon cancer cell lines, activated macrophages and monocytes.

PPARγ is also involved in lipid metabolism, adipocyte differentiation, and glucose homeostasis and insulin sensitivity. A less-credible positive association also reported one variant (rs2066847) in NOD2 gene; however, this association was not replicated. Evidence from experimental studies in mice investigating the role of Nod2 in colorectal tumour risk has also been inconsistent. The genetic loci however, we would like to highlight a number of less-substantial heterogeneity or the high possibility of false positive; 83 variants were classified as less-credible 'ants were classified as less-credible associations with 63 variants of 52 genes have been reported to be associated with telomere length and the risk of different types of cancer and chronic diseases other than cancer. For the association with the rs16892766 variant at 8q23.3 tagging EIF3H, another variant rs1688589 was previously reported with the lowest p value for the association with CRC in this locus. The analysis clearly identifies 14 variants at 12 loci for which there is robust evidence of their impact on CRC risk, 63 variants at 52 loci for which further evidence through international collaboration should be generated and 231 variants for which the overall evidence does not support any association with CRC risk. With increasing availability of data from multiple SNPs, it is clear that studies to test associations must achieve very high levels of statistical stringency. Nonetheless, the analysis here provides a resource for mining available data and puts into context the sample sizes required for the identification of true associations for common genetic variants. Future resequencing studies are expected to identify rarer variants (eg, prevalence 0.05%–5%) with intermediate or perhaps even large effects, and GWAS of structural variation will likely identify deletions, amplifications and other copy number variations that may also influence CRC risk. This study highlights a number of common genetic variants that could be incorporated into genetic risk-prediction algorithms as further risk factors are identified and highlights the loci at which further research effort should be targeted. All data are available from the CRCgene2 database.

**Author affiliations**

School of Epidemiology and Public Health, Faculty of Medicine, University of Ottawa, Ottawa, Ontario, Canada

Centre for Global Health, Usher Institute, The University of Edinburgh, Edinburgh, UK

Department of Epidemiology, College of Preventive Medicine, Third Military Medical University, Chongqing, Chongqing, China

Colón Cancer Genetics Group, Cancer Research UK Edinburgh Centre and Medical Research Council Human Genetics Unit, Medical Research Council Institute of Genetics & Molecular Medicine, The University of Edinburgh, Edinburgh, UK

Jockey Club School of Public Health and Primary Care, The Chinese University of Hong Kong, Shenzhen, Hong Kong

Gastroenterology Department, Institut D’Investigacions Biomèdiques August Pi i Sunyer (IDIBAPS), Hospital Clinic de Barcelona, Universitat de Barcelona, Centro de Investigación Biomédica en Red de Enfermedades Hepáticas y Digestivas (CIBERHD), Barcelona, Spain

Fundación Pública Galéga de Medicina Xénica, Grupo de Medicina Xénica, Santiago de Compostela, Spain; Instituto de Investigación Sanitaria de Santiago
REFERENCES


36 Prokunina-Olsson L, Hall JL. No effect of cancer-associated SNP rs6983267 in the 8q24 region on co-expression of myc and TC7FL2 in normal colon tissue. *Mol Cancer* 2009;8:96.


