Genetic architectures of proximal and distal colorectal cancer are partly distinct
Jeroen R Huyghe, Tabitha A Harrison, Stephanie A Bien, Heather Hampel, Jane C Figueiredo, Stephanie L Schmit, David V Conti, Sai Chen, Conghui Qu, Yi Lin, et al.

To cite this version:
Jeroen R Huyghe, Tabitha A Harrison, Stephanie A Bien, Heather Hampel, Jane C Figueiredo, et al.. Genetic architectures of proximal and distal colorectal cancer are partly distinct. Gut, 2021, 70 (7), pp.1325 - 1334. 10.1136/gutjnl-2020-321534 . hal-03915488

HAL Id: hal-03915488
https://hal.science/hal-03915488
Submitted on 29 Dec 2022

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L’archive ouverte pluridisciplinaire HAL, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d’enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.
Genetic architectures of proximal and distal colorectal cancer are partly distinct


ABSTRACT

Objective An understanding of the etiologic heterogeneity of colorectal cancer (CRC) is critical for improving precision prevention, including individualized screening recommendations and the discovery of novel drug targets and repurposable drug candidates for

© Author(s) (or their employer(s)) 2021. Re-use permitted under CC BY-NC. No commercial re-use. See rights and permissions. Published by BMJ.
**INTRODUCTION**

Despite improvements in prevention, screening and therapy, colorectal cancer (CRC) remains one of the leading causes of cancer-related death worldwide, with an estimated 53 200 fatal cases in 2020 in the USA alone.\(^1\) CRCs that arise proximal (right) or distal (left) to the splenic flexure differ in age-specific and sex-specific incidence rates, clinical, pathological and tumour molecular features.\(^2,3\) These observed differences reflect a complex interplay between differential exposure of colorectal crypt cells to local environmental carcinogenic and protective factors in the luminal content (including the microbiome), and distinct inherent biological characteristics that may influence neoplasia risk, including sex and differences between anatomical segments in embryonic origin, development, physiology, function and mucosal immunity. The precise extrinsic and intrinsic aetiological factors involved, their relative contributions, and how they interact to influence the carcinogenic process remain largely elusive.

An individual’s genetic background plays an important role in the initiation and development of CRC. Based on twin registries, heritability is estimated to be around 35%.\(^6\) Since genome-wide association studies (GWASs) became possible just over a decade ago, over 100 independent common genetic variant associations for overall CRC risk have been identified, over half of which were identified in the past few years.\(^7,8\) Three decades ago, based on observed similarities between Lynch syndrome and proximal CRC, and between familial adenomatous polyposis and distal CRC, Bufill proposed the existence of two distinct genetic categories of CRC according to the location of the primary tumour.\(^2\) However, given that genetic variants that influence CRC risk typically have small effect sizes, until very recently, sample sizes did not provide adequate statistical power to conduct meaningful subsite analyses. As a consequence, GWASs to detect genetic associations specific to CRC case subgroups defined by primary tumour anatomic subsite have not been reported yet. Similarly, a comprehensive analysis of the extent to which allelic risk of known GWAS-identified variants differs by primary tumour anatomic subsite is lacking.

To address the major gap in our knowledge of the differential role that genetic variants, genes and pathways play in mechanisms of proximal and distal CRC carcinogenesis, we analysed clinical and genome-wide genotype data for 112 373 CRC cases and controls. First, to discover new loci and genetic risk variants with site-specific allelic effects, we conducted GWASs of case subgroups defined by the location of their primary tumour within the colorectum. Next, we systematically characterised heterogeneity of allelic effects between primary tumour subsites for new and previously identified CRC risk variants to identify loci with shared and site-specific allelic effects.

---

**Significance of this study**

**What is already known on this subject?**

- Heterogeneity among colorectal cancer (CRC) tumours originating at different locations of the colorectum has been revealed in somatic genomes, epigenomes and transcriptomes, and in some established environmental risk factors for CRC.
- Genome-wide association studies (GWASs) have identified over 100 genetic variants for overall CRC risk; however, a comprehensive analysis of the extent to which genetic risk factors differ by the anatomical sublocation of the primary tumour is lacking.

**What are the new findings?**

- In this large consortium-based study, we analysed clinical and genome-wide genotype data of 112 373 CRC cases and controls of European ancestry to comprehensively examine whether CRC case subgroups defined by anatomical sublocation have distinct germline genetic aetiologies.
- We discovered 13 new loci at genome-wide significance (p<5x10^{-8}) that were specific to certain anatomical sublocations and that were not reported by previous GWASs for overall CRC risk; multiple lines of evidence support strong candidate target genes at several of these loci, including PTGER3, LCT, MLH1, CDX1, KLF14, PYGL, BCL11B and BMP7.
- Systematic heterogeneity analysis of genetic risk variants for CRC identified thus far, revealed that genetic architectures of proximal and distal CRC are partly distinct, and demonstrated that distal colon and rectal cancer have very similar germline genetic aetiologies.
- Taken together, our results further support the idea that tumours arising in different anatomical sublocations of the colorectum may have distinct aetiologies.

**How might it impact on clinical practice in the foreseeable future?**

- Our results provide an informative resource for understanding the differential role that genetic variants, genes and pathways may play in the mechanisms of proximal and distal CRC carcinogenesis.
- The new insights into the aetiologies of proximal and distal CRC may inform the development of new precision prevention strategies, including individualised screening recommendations and the discovery of novel drug targets and repurposable drug candidates for chemoprevention.
- Our findings suggest that future studies of aetiological risk factors for CRC and molecular mechanisms of carcinogenesis should take into consideration the anatomical sublocation of the colorectal tumour. In particular, our results argue against lumping proximal and distal colon cancer cases.

---

chemoprevention. Known differences in molecular characteristics and environmental risk factors among tumors arising in different locations of the colorectum suggest partly distinct mechanisms of carcinogenesis. The extent to which the contribution of inherited genetic risk factors for CRC differs by anatomical subsite of the primary tumour has not been examined.

**Design** To identify new anatomical subsite-specific risk loci, we performed genome-wide association study (GWAS) meta-analyses including data of 48 214 CRC cases and 64 159 controls of European ancestry. We characterised effect heterogeneity at CRC risk loci using multinomial modelling.
METHODS
Detailed methods are provided in online supplemental materials.

Samples and genotypes
This study included clinical and genotype data for 48,214 CRC cases and 64,159 controls from three consortia: Genetics and Epidemiology of Colorectal Cancer Consortium (GECCO), Colorectal Cancer Transdisciplinary Study (CORECT) and Colorectal Cancer Family Registry (CCFR). Online supplemental table 1 provides details on sample numbers and demographic characteristics by study. All study participants were of genetically inferred European-ancestry. Across studies, participant recruitment occurred between the early 1990s and the 2010s. Details of genotype data sets, genotype QC, sample selection and studies included in this analysis have been published previously. All participants provided written informed consent, and each study was approved by the relevant research ethics committee or institutional review board.

Colon tumour anatomic sublocation definitions
We defined proximal colon cancer as any primary tumour arising in the cecum, ascending colon, hepatic flexure or transverse colon; distal colon cancer as any primary tumour arising in the splenic flexure, descending colon or sigmoid colon; and rectal cancer as any primary tumour arising in the rectum or recto-sigmoid junction. For the GWAS discovery analyses, we analysed five case subgroups based on primary tumour sublocation. In addition to the three aforementioned mutually exclusive case sets (proximal colon, distal colon and rectal cancer), we defined colon cancer and distal/left-sided colorectal cancer case sets. Colon cancer cases comprised combined proximal colon and distal colon cancer cases, and additional colon cases with unspecified site. In the distal/left-sided colorectal cancer cases analysis, we combined distal colon and rectal cancer cases based on the different embryonic origins of the proximal colon versus the distal colon and rectum. Online supplemental figure 1 and table 1 summarise distributions of age of diagnosis by sex and primary tumour site.

Statistical analysis
GWAS meta-analyses
We imputed all genotype datasets to the Haplotype Reference Consortium panel. In brief, we phased all genotyping array data sets using SHAPEIT and used the Michigan Imputation Server for imputation. Within each dataset, variants with an imputation accuracy $r^2 \geq 0.3$ and minor allele count $\geq 50$ were tested for association with CRC case subgroups. Variants that only passed filters in a single dataset were excluded. We assumed an additive model using imputed genotype dosage in a logistic regression adjusted for age, sex and study or genotyping project-specific covariates, including principal components to adjust for population structure. Details of covariate corrections have been published previously. Because Wald tests can be anticonservative for rare variants, we performed likelihood ratio tests and combined association summary statistics across sample sets via fixed-effects meta-analysis employing Stouffer’s method, implemented in the METAL software. Reported $p$ values are based on this analysis. Reported combined OR estimates and 95% CIs are based on an inverse variance-weighted fixed-effects meta-analysis.

Heterogeneity in allelic effect sizes between tumour anatomic sublocations
To characterise tumour subsite-specificity and effect size heterogeneity across tumour subsites for new loci, and for established loci for overall CRC, we examined association evidence in three different ways. First, for each index variant we created forest plots of OR estimates from GWAS meta-analyses for proximal colon, distal colon and rectal cancer. Second, we tested for heterogeneity using multinomial logistic regression. In brief, after pooling of datasets, we performed a likelihood ratio test comparing a model in which ORs for the risk variant were allowed to vary across tumour subtypes, to a model in which ORs were constrained to be the same across tumour subtypes. Third, inspired by reference, we used a multinomial logistic regression-based model selection approach to assess which configuration of tumour subtypes is most likely to be associated with a given variant. For each variant, we defined and fitted 11 possible causal risk models specifying variant effect configurations that vary or are constrained to be equal among subsets of tumour subtypes (online supplemental table 2). We then identified and report the best fitting model using the Bayesian information criterion (BIC). For each model $i$ we calculated $\Delta \text{BIC} = \text{BIC}_i - \text{BIC}_{min}$, where $\text{BIC}_{min}$ is the BIC value for the best model. Models with $\Delta \text{BIC} \leq 2$ were considered to have substantial support and indistinguishable from the best model. For these variants, we do not report a single best model. Analyses were carried out using the VGAM R package. The list of index variants for previously published CRC risk signals is based on Huyghe et al.

Pathway enrichment analyses
We used the Pascal programme to compute pathway enrichment score $p$ values from genome-wide summary statistics. The gene set library used comprises the combined KEGG, REACTOME and BIOCARTE databases.

Genomic annotation of new GWAS loci and gene prioritisation
We annotated all new loci with five types of functional and regulatory genomic annotations: (i) cell-type-specific regulatory annotations for histone modifications and open chromatin, (ii) nonsynonymous coding variation, (iii) evidence of transcription factor binding, (iv) predicted functional impact across different databases, (v) colocalisation with expression quantitative trait loci (eQTL) signals. Genes were further prioritised based on biological relevance, colorectal tissue expression, presence of associated non-synonymous variants predicted to be deleterious, evidence from functional studies, somatic alterations or familial syndromes. Details are in online supplemental materials.

RESULTS
The final analyses included data for 48,214 CRC cases and 64,159 controls of European ancestry. To discover new loci and genetic risk variants with site-specific allelic effects, we conducted five genome-wide association scans of case subsets defined by the location of their primary tumour within the colorectum: proximal colon cancer ($n=15,706$), distal colon cancer ($n=14,376$), rectal cancer ($n=16,212$), colon cancer, in which we omitted rectal cancer cases ($n=32,002$), and distal/left-sided CRC, in which we combined distal colon and rectal cancer cases ($n=30,588$). Next, we systematically characterised heterogeneity of allelic effects between tumour subtypes for new and previously identified CRC risk variants to identify loci with shared and site-specific allelic effects.
New colorectal cancer risk loci

Across the five CRC case subgroup GWAS meta-analyses, a total of 11 947 015 single nucleotide variants (SNVs) were analysed. Inspection of genomic control inflation factors and quantile-quantile plots of test statistics indicated no residual population stratification issues (online supplemental materials and figure 2). Across tumour subtypes, we identified 13 loci that mapped outside regions previously implicated by GWAs for overall CRC risk (closest known locus 3.1 megabases away) and that reached genome-wide significance ($p<5 \times 10^{-8}$) in at least one of the meta-analyses (table 1, figure 1, online supplemental figures 3 and 4). Seven of the new loci passed a Bonferroni-adjusted genome-wide significance threshold correcting for five case subgroups analysed (table 1). All lead variants were well imputed (minimum average imputation $r^2=0.788$), had minor allele frequency (MAF) >1%, and displayed no significant heterogeneity between sample sets (Cochran’s Q heterogeneity test $p>0.05$; table 1).

The novel associations showing the strongest statistical evidence were obtained for proximal colon cancer and mapped near MLH1 on 3p22.2 (rs1800734, $p=3.8 \times 10^{-15}$) and near BCL11B on 14q32.2 (rs80158569, $p=8.6 \times 10^{-11}$). These loci showed strongly proximal cancer-specific associations. The proximal colon analysis also yielded a locus on 14q32.12 (rs61975764, $p=2.8 \times 10^{-8}$) that showed attenuated effects for other tumour subtypes (figure 1 and online supplemental table 3). Most new loci (six) were discovered in the left-sided CRC analysis: 2q21.3 (rs1446585, $p=3.3 \times 10^{-8}$), near CDX1 on 5q32 (rs2302274, $p=4.9 \times 10^{-8}$), near KLF14 on 7q32.3 (rs73161913, $p=1.3 \times 10^{-7}$), 10q23.31 (rs7071258, $p=8.4 \times 10^{-7}$), 19p13.13 (rs62131228, $p=2.4 \times 10^{-8}$) and near BMP7 on 20q13.31 (rs6014965, $p=4.5 \times 10^{-8}$). The rectal cancer analysis identified an additional locus near PYGL on 14q22.1 (rs28611105, $p=4.7 \times 10^{-7}$) that showed an attenuated effect for distal colon cancer (figure 1 and online supplemental table 3). No additional new loci were detected in the distal colon analysis. The colon cancer analysis identified three new loci: near PTGER3 on 1p31.1 (rs3124454, $p=1.4 \times 10^{-8}$), 3p21.2 (rs353548, $p=1.3 \times 10^{-8}$) and 22q13.31 (rs736037, $p=2.8 \times 10^{-8}$).

Genomic annotations and most likely target gene(s) at new loci

To gain insight into molecular mechanisms underlying new association signals, and to identify candidate causal variants and target gene(s), we annotated signals with functional and regulatory genomic annotations, assessed colocalisation with eQTLs, and performed literature-based gene prioritisation. Results for all new signals are given in online supplemental tables 4 and 5, and candidate target genes are also given in table 1. Notable and strong candidate target genes include PTGER3, LCT, MLH1, CDX1, KLF14, PYGL, RIN3, BCL11B and BMP7. Strong candidate causal variants were identified at loci 2q21.3 (rs4988235; LCT), 3p22.2 (rs1800734; MLH1), 14q32.12 (rs61975764; RIN3) and 14q32.3 (rs80158569; BCL11B). A detailed interpretation of candidate causal variants and target genes is deferred to the Discussion section.

Risk heterogeneity between tumour anatomical sublocations

Multinomial logistic regression modelling of 96 known and 13 newly identified risk variants showed the presence of substantial risk heterogeneity between cancer in the proximal colon, distal colon and rectum. For 61 variants, the heterogeneity $p$ value ($p_{\text{het}}$) was not significant ($p_{\text{het}}>0.05$). For 51 of those variants,

| Table 1 New genome-wide significant colorectal cancer risk loci identified by genome-wide association analysis of case subgroups defined by primary tumour anatomic subsite |
|----------------------------------|-----------------|----------|-------------|---------------|----------------|----------|------|------|
| Tumour site* | Locus | Nearby gene(s) | rsID lead variant | Chr. | Position (base pair) | Alleles (risk/other) | p value | OR 95% CI |
| Left-sided 2q21.3 | MLH1 | PYGL | rs1446585 | 2 | 7 049 448 G | 3.81 10^{-8} | 1 | 1.07 1.04 to 1.09 | 4.71 10^{-9} |
| Left-sided 5q32 | CDX1 | KLF14 | rs2302274 | 5 | 14 954 438 G | 2 15 10^{-8} | 1 | 1.08 1.05 to 1.11 | 8.41 10^{-8} |
| Left-sided 7q32.3 | MLH1 | KLF14;LINC00513 | rs28611105 | 7 | 13 067 779 G | 2 15 10^{-8} | 1 | 1.08 1.05 to 1.11 | 4.61 10^{-9} |
| Left-sided 10q23.31 | MLH1 | BMP7 | rs7071258 | 10 | 91 546 846 G | 2 15 10^{-8} | 1 | 1.08 1.05 to 1.11 | 8.41 10^{-8} |
| Left-sided 19p13.3 | MLH1 | STK11; SBNO2 | rs6014965 | 19 | 5 139 648 G | 2 15 10^{-8} | 1 | 1.08 1.05 to 1.11 | 4.61 10^{-9} |
| Left-sided 20q13.31 | MLH1 | BMP7 | rs6014965 | 20 | 5 552 138 G | 2 15 10^{-8} | 1 | 1.08 1.05 to 1.11 | 4.61 10^{-9} |
| Rectal 14q22.1 | PTGER3 | PYGL;NIN; ABHD12B | rs73161913 | 14 | 93 453 548 G | 2 15 10^{-8} | 1 | 1.08 1.05 to 1.11 | 8.41 10^{-8} |
| Rectal 14q32.12 | MLH1 | KLF14 | rs61975764 | 14 | 93 453 548 G | 2 15 10^{-8} | 1 | 1.08 1.05 to 1.11 | 8.41 10^{-8} |
| Rectal 20q13.31 | MLH1 | BMP7 | rs6014965 | 20 | 5 552 138 G | 2 15 10^{-8} | 1 | 1.08 1.05 to 1.11 | 4.61 10^{-9} |
| Rectal 22q13.31 | MLH1 | BMP7 | rs6014965 | 22 | 5 552 138 G | 2 15 10^{-8} | 1 | 1.08 1.05 to 1.11 | 4.61 10^{-9} |

*Colon: proximal colon+distal colon+colon, unspecified site; left-sided 2q21.3; right-sided 5q32; left-sided 7q32.3; right-sided 10q23.31; left-sided 19p13.3; right-sided 20q13.31. Alleles are on the + strand. All p values reported in this table are from a sample size-adjusted genome-wide significance (5E-08/5=1E-08), corrected for the number of CRC case subgroups analysed. RAF, risk allele frequency.
a multinomial model in which ORs were identical for the three cancer sites provided the best fit, and for 8 of the remaining 10 variants, this model did not significantly differ from the best fitting model (online supplemental tables 2, 3 and 7; figure 5).

Among the 109 known or new variants, 48 showed at least some evidence of heterogeneity with $p_{het} < 0.05$, and after Holm-Bonferroni correction for multiple testing, 14 variants showed strong evidence of heterogeneity remained significant ($p_{het} < 4.6 \times 10^{-4}$). These included 10 variants previously reported in GWAS for overall CRC risk.

For 17 out of the 48 variants with $p_{het} < 0.05$, the best-fitting model supported an effect limited to left-sided CRC (figure 2 and online supplemental tables 3 and 7). Of these 17 variants, 6 were in the list of variants with the strongest evidence of heterogeneity ($p_{het} < 4.6 \times 10^{-4}$), including the following previously reported loci: C11orf53-COLCA1-COLCA2 on 11q23.1 ($p_{het}=6.0 \times 10^{-18}$), APC on 5q22.2 ($p_{het}=2.3 \times 10^{-10}$), GATA3 on 10p14 ($p_{het}=1.7 \times 10^{-9}$), CTNNB1 on 3p22.1 ($p_{het}=9.8 \times 10^{-8}$), Rab40B-METTL1 on 17q25.3 ($p_{het}=3.6 \times 10^{-8}$), and CDKN1A on 6p21.2 ($p_{het}=1.6 \times 10^{-7}$). Inspection of forest plots and association evidence also suggest stronger risk effects for left-sided tumours for the following additional five known loci: TET2 on 4q24, VTI1A on 10q25.2, two independent signals near POLD3 on 11q13.4, and BMP4 on 14q22.2.

For 5 out of the 49 variants with $p_{het} < 0.05$, a model with association with rectal cancer risk, but no association with rectal cancer risk, provided the best fit (online supplemental tables 3 and 7). These involve the following loci: PTGER3 on 1p31.1, STAB1-TLR9 on 3p21.2, HLA-B-MICA/NFKBIL1-TNF on 6p21.33, NOS1 on 12q24.22 and LINCO0673 on 17q24.3. Association evidence also suggests stronger risk effects for colon tumours for one of two independent signals near PTPN1 on 20q13.13.

Evidence from the three approaches (figure 1; online supplemental tables 3 and 7) indicates that only two loci are strongly proximal colon cancer-specific: MLH1 on 22p2.2 ($p_{het}=5.4 \times 10^{-10}$), and BCL11B ($p_{het}=1.5 \times 10^{-5}$) on 14q32.2. Finally, for only one variant, at one of two independent loci near SATB2 on 2q33.1, a model with a rectal cancer-specific association provided the best fit, but association evidence shows attenuated effects for proximal and distal colon cancer. OR estimates also suggest stronger risk effects for rectal cancer at the known

---

**Figure 1** Primary tumour site-specific associations for the lead single nucleotide polymorphisms (SNPs) of the 13 colorectal cancer risk loci not reported in previous genome-wide association studies. The forest plot shows the (log-additive) OR estimates together with 95% CIs. For clarity, this figure only shows results for the proximal colon, distal colon and rectal cancer case subgroup analyses.

**Figure 2** Loci showing association with risk of distal colorectal cancer (i.e. distal colon+rectal), but attenuated or no evidence for association with proximal colon cancer risk. The forest plot shows the (log-additive) OR estimates for the lead single nucleotide polymorphisms (SNPs) at the loci, together with 95% CIs, from the genome-wide association study meta-analyses of case subgroups defined by primary tumour anatomical subsite for proximal colon, distal colon and rectal. Best model is the best-fitting multinomial logistic regression model according to the Bayesian information criterion (BIC). Models are defined in online supplemental table 2. $P_{het}$ is the $p$ value from a test for heterogeneity of allelic effects across tumour subites.
lnc LAMC1 on 1q25.3, and CTNNB1 on 3p22.1, and at new locus PYGL on 14q22.1.

Pathway enrichment analyses
To explore whether biological pathways play different roles in tumourigenesis of proximal and distal CRC, we conducted pathway enrichment analyses of GWAS summary statistics. There was no clear and strong evidence for differential involvement of pathways; pathways that were Bonferroni-significant for one anatomical subtype, reached at least suggestive significance levels for other subtypes (online supplemental table 8). Several of the Bonferroni-significant pathways related to transforming growth factor β (TGFβ) signalling.

DISCUSSION
It has long been recognised that CRCs arising in different anatomical segments of the colorectum differ in age-specific and sex-specific incidence rates, clinical, pathological and tumour molecular features. However, our understanding of the aetiological factors underlying these medically important differences has remained scarce. This study aimed to examine whether the contribution of common germline genetic variants to CRC carcinogenesis differs by anatomical sublocation. The large sample size comprising 112 373 cases and controls provided robustness to 1.14 (95% CI 1.09 to 1.19). Because this region has been under strong selection, there was no clear and strong evidence for differential involvement of pathways; pathways that were Bonferroni-significant for one anatomical subtype, reached at least suggestive significance levels for other subtypes (online supplemental table 8). Several of the Bonferroni-significant pathways related to transforming growth factor β (TGFβ) signalling.

At the MLH1 gene promoter region on 3p22.2, associated with left-sided CRC, the intestine-specific transcription factor caudal-type homeobox 1 (CDX1) encodes a key regulator of differentiation of enterocytes in the normal intestine and of CRC cells. CDX1 is central to the capacity of colon cells to differentiate and promotes differentiation by repressing the polycomb complex protein BMI1 which promotes stemness and self-renewal. The repression of BMI1 is mediated by microRNA-215 which acts as a target of CDX1 to promote differentiation and inhibit stemness.

In a region of extensive LD on locus 2q21.1, lead SNP rs1446585, associated with left-sided CRC, is in strong LD with functional SNP rs4988235 (LD r2=0.854) in the cis-regulatory element of the lactase (LCT) gene. In Europeans, the rs4988235 genotype determines the lactase persistence phenotype, or the ability to digest lactose in adulthood. The p value for functional SNP rs4988235 under an additive model was 7.0×10^{-7}. The allele determining lactase persistence (T) is associated with decreased CRC risk. This is consistent with a previously reported association between low lactase activity defined by the CC genotype and CRC risk in the Finnish population.

The protective effect conferred by the lactase persistence genotype is likely mediated by dairy products and calcium which are known protective factors for CRC. When we tested for association with left-sided CRC assuming a dominant model, associations for rs1446585 and rs4988235 became more significant with p values of 4.4×10^{-11} and 1.4×10^{-10}, respectively. For functional SNP rs4988235, the OR estimate for having genotype CC versus CT or TT, and left-sided CRC was 1.14 (95% CI 1.09 to 1.19). Because this region has been under strong selection, it is particularly prone to population stratification. However, we adjusted for genotype principal components, and the association showed a consistent direction of effect across sample sets (online supplemental table 6), suggesting this association is not spurious.

Candidate genes at left-sided CRC loci 7q32.2 and 2q13.31 are involved in TGFβ signalling. At 7q32.3, gene Krüppel-like factor 14 (KLF14) is a strong candidate. We previously reported loci at known CRC oncogene KLF5 and at KLF2. The imprinted gene KLF14 shows monoallelic maternal expression, and is induced by TGFβ to transcriptionally corepress the TGFβ receptor 2 (TGFBR2) gene. A cis-eQTL for KLF14, uncorrelated with our lead SNP rs73161913, acts as a master regulator related to multiple metabolic phenotypes, and a nearby independent variant is associated to basal cell carcinoma. For both reported associations, effects depended on parent-of-origin of risk alleles. The association with metabolic phenotypes also depended on sex. We did not find evidence for strong sex-dependent effects (men: OR=1.13, 95% CI 1.07 to 1.20; women: OR=1.17, 95% CI 1.09 to 1.25). Further investigation is warranted to analyse parent-of-origin effects. At 20q13.31, gene bone morphogenetic protein 7 (BMP7) is a strong candidate. BMP7 signalling in TGFBR2-deficient stromal cells promotes epithelial carcinogenesis through SMAD4-mediated signalling.
BMP7 expression correlates with parameters of pathological aggressiveness such as liver metastasis and poor prognosis.46

On 1q42.21, the single locus identified only in the rectal cancer analysis, GTex data show that, in gastrointestinal tissues, lead SNP rs28611105 colocalises with a cis-eQTL coregulating expression of genes PYGL, ABHD12B and NIN. We reported an association between genetically predicted glycosylphosphatidylinositol (GPI) expression and CRC risk in a transcriptome-wide association study.47 This glycosylation metabolism gene plays an important role in sustaining proliferation and preventing premature senescence in hypoxic cancer cells.48

At 1p31.1, identified in the colon cancer analysis, PTGER3 encodes prostaglandin E receptor 3, a receptor for prostaglandin E2 (PGF2), a potent pro-inflammatory metabolite biosynthesised by cyclooxygenase-2 (COX-2). COX-2 plays a critical role in mediating inflammatory responses that lead to epithelial malignancies. The anti-inflammatory activity of non-steroidal anti-inflammatory drugs (NSAIDs) such as aspirin and ibuprofen operates mainly through COX-2 inhibition, and long-term NSAID use decreases CRC incidence and mortality.49 PGF2 is required for the activation of β-catenin by Wnt in stem cells,50 and promotes colon cancer cell growth.51 PTGER3 plays an important role in suppression of cell growth and its downregulation was shown to enhance colon carcinogenesis.52

Previous CRC GWASs had already reported allelic effect heterogeneity between tumour sites, including for 10p14, 11q23 and 18q21 but only contrasted colon and rectal tumours, without distinguishing between proximal and distal colon.53-54 Sample size and timing of the present study enabled systematic characterisation of allelic effect heterogeneity between more refined tumour anatomical sublocalizations, and for a much expanded catalogue of risk variants. Our analysis revealed substantial, previously unappreciated allelic effect heterogeneity between proximal and distal CRC. Results further show that distal colon and rectal cancer have very similar germline genetic aetiologies. Our findings at several loci are consistent with CRC tumour molecular studies. Consensus molecular subtypes (CMSs), which are based on tumour gene expression, are differentially distributed between proximal and distal CRCs. The canonical CMS (CMS2) is enriched in distal CRC (56% vs 26% for proximal CRC) and is characterised by upregulation of Wnt downstream targets.55

We found that variant associations near Wnt/β-catenin pathway genes APC and CTNNB1 were confined to distal CRC. We also found that associations for variants near genes BOC and FOXL1, members of the Hedgehog signalling pathway, were confined to distal CRC. Suggesting that Wnt and Hedgehog signalling may contribute more to the development of distal CRC tumours. However, pathway enrichment analyses did not provide clear evidence for differential involvement of pathways, suggesting perhaps that associations for proximal and distal CRC mostly converge on the same pathways. Pathway analysis results should, however, be interpreted taking into consideration the limitations of available approaches. Genetic variants were mapped to the nearest gene which is often not the target gene.

The precise intrinsic or extrinsic effect modifiers explaining observed allelic effect heterogeneity between anatomical subsites remain unknown and further research is needed. Short-chain fatty acids, in particular butyrate, produced by microbiota through fermentation of dietary fibre in the colon may be involved. Concentrations of butyrate, which plays a multifaceted antitumorigenic role in maintaining gut homeostasis, are much higher in proximal colon.56 Moreover, the known chemopreventive role of butyrate may involve modulation of signalling pathways including TGFβ and Wnt.57 This may contribute to possible differences between anatomical segments in colorectal crypt cellular dynamics.

One limitation of our study is that we have not performed GWAS analyses of case subgroups based on more detailed anatomical sublocalizations. However, given current sample size, such analyses would result in reduced statistical power owing to reduced sample sizes and the aggravated multiple testing burden. As another limitation, our study was based on European-ancestry subjects and it remains to be determined whether findings are generalisable to other ancestries.

In conclusion, germline genetic data support the idea that proximal and distal colorectal cancer have partly distinct aetiologies. Our results further demonstrate that distal colon and rectal cancer have very similar germline genetic aetiologies and argue against lumping proximal and distal colon cancer in studies of aetiological factors. Future genetic studies should take into consideration differences between primary tumour anatomical subsites. A better understanding of differing carcinogenic mechanisms and neoplastic transformation risk in proximal and distal colorectum can inform the development of novel precision treatment and prevention strategies through the discovery of novel drug targets and repurposable drug candidates for treatment and chemoprevention, and improved individualised screening recommendations based on risk prediction models incorporating tumour anatomical subsite.

Author affiliations
1Public Health Sciences Division, Fred Hutchinson Cancer Research Center, Seattle, Washington, USA
2Division of Human Genetics, Department of Internal Medicine, The Ohio State University Comprehensive Cancer Center, Columbus, Ohio, USA
3Department of Preventive Medicine, Keck School of Medicine of the University of Southern California, Los Angeles, California, USA
4Department of Medicine, Samuel Oschin Comprehensive Cancer Institute, Cedars-Sinai Medical Center, Los Angeles, California, USA
5Genomic Medicine Institute, Cleveland Clinic, Cleveland, Ohio, USA
6Department of Preventive Medicine and USC Norris Comprehensive Cancer Center, Keck School of Medicine of the University of Southern California, Los Angeles, California, USA
7Department of Biostatistics and Center for Statistical Genetics, University of Michigan, Ann Arbor, Michigan, USA
8Department of Medicine, University of North Carolina School of Medicine, Chapel Hill, North Carolina, USA
9Department of Epidemiology and Biostatistics, Imperial College London, London, UK
10Department of Human Genetics, Graduate School of Public Health, University of Pittsburgh, Pittsburgh, Pennsylvania, USA
11UPMC Hillman Cancer Center, Pittsburgh, Pennsylvania, USA
12Translational Genomics Research Institute - An Affiliate of City of Hope, Phoenix, Arizona, USA
13Department of Radiation Sciences, Oncology Unit, Umeå University, Umeå, Sweden
14Public Health Division of Gipuzkoa, Health Department of Basque Country, San Sebastian, Spain
15Department of Biostatistics, University of Washington, Seattle, Washington, USA
16Laboratoire de Mathématiques Appliquées MAPS (UMR CNRS 8145), Université Paris Descartes, Paris, France
17Centre for Research in Epidemiology and Population Health (CESP), Institut pour la Santé et la Recherche Médicale (INSERM) U1018, Université Paris-Saclay, Villejuif, France
18Cancer Epidemiology Unit, Nuffield Department of Population Health, University of Oxford, Oxford, UK
19Division de Recherche, Kaiser Permanente Northern California, Oakland, California, USA
20Division of Population and Quantitative Health Sciences, Case Western Reserve University, Cleveland, Ohio, USA
21Department of Internal Medicine, University of Utah Health, Salt Lake City, Utah, USA
22Departments of Cancer Biology and Genetics and Internal Medicine, The Ohio State University, Columbus, Ohio, USA
23Division of Human Nutrition and Health, Wageningen University & Research, Wageningen, The Netherlands
24Unit of Nutrition and Cancer, Cancer Epidemiology Research Program, Catalan Institute of Oncology - IDIBELL, L'Hospital de Llobregat, Barcelona, Spain

REFERENCES


Daniel D Buchanan http://orcid.org/0000-0003-2225-6675
Andrew T Chan http://orcid.org/0000-0001-7284-6767
Andrea Gsur http://orcid.org/0000-0002-9795-1528
Jochen Hampe http://orcid.org/0000-0002-2421-6127
Michael Hoffmeister http://orcid.org/0000-0002-8307-3197
Wen-Yi Huang http://orcid.org/0000-0002-8300-3368
Mazda Janef http://orcid.org/0000-0002-0573-1852
Neil Murphy http://orcid.org/0000-0003-3347-8249
Ulrice Peters http://orcid.org/0000-0001-5666-9318
57 McIntabney SM, Heneghan TM. Short chain fatty acids in the colon and peripheral tissues: a focus on butyrate, colon cancer, obesity and insulin resistance. Nutrients 2017;9:9.