



OPEN ACCESS

Original research

Genetic architectures of proximal and distal colorectal cancer are partly distinct

Jeroen R Huyghe ,¹ Tabitha A Harrison,¹ Stephanie A Bien,¹ Heather Hampel,² Jane C Figueiredo,^{3,4} Stephanie L Schmit,⁵ David V Conti,⁶ Sai Chen,⁷ Conghui Qu,¹ Yi Lin,¹ Richard Barfield,¹ John A Baron,⁸ Amanda J Cross,⁹ Brenda Diergaarde,^{10,11} David Duggan,¹² Sophia Harlid,¹³ Liher Imaz,¹⁴ Hyun Min Kang,⁷ David M Levine,¹⁵ Vittorio Perduca,^{16,17} Aurora Perez-Cornago,¹⁸ Lori C Sakoda,^{1,19} Fredrick R Schumacher,²⁰ Martha L Slattey,²¹ Amanda E Toland,²² Fränzel J B van Duijnhoven,²³ Bethany Van Guelpen,¹³ Antonio Agudo,²⁴ Demetrius Albanes,²⁵ M Henar Alonso,^{26,27,28} Kristin Anderson,²⁹ Coral Arnau-Collell,³⁰ Volker Arndt,³¹ Barbara L Banbury,¹ Michael C Bassik,³² Sonja I Berndt,²⁵ Stéphane Bézieau,³³ D Timothy Bishop,³⁴ Juergen Boehm,³⁵ Heiner Boeing,³⁶ Marie-Christine Boutron-Ruault,^{17,37} Hermann Brenner ,^{31,38,39} Stefanie Brezina,⁴⁰ Stephan Buch,⁴¹ Daniel D Buchanan ,^{42,43,44} Andrea Burnett-Hartman,⁴⁵ Bette J Caan,⁴⁶ Peter T Campbell,⁴⁷ Prudence R Carr,⁴⁸ Antoni Castells,³⁰ Sergi Castellví-Bel,³⁰ Andrew T Chan ,^{49,50,51,52,53,54} Jenny Chang-Claude,^{55,56} Stephen J Chanock,²⁵ Keith R Curtis,¹ Albert de la Chapelle,⁵⁷ Douglas F Easton,⁵⁸ Dallas R English,^{42,59} Edith J M Feskens,²³ Manish Gala,^{49,51} Steven J Gallinger,⁶⁰ W James Gauderman,⁶ Graham G Giles,^{42,59} Phyllis J Goodman,⁶¹ William M Grady,^{62,63} John S Grove,⁶⁴ Andrea Gsur ,⁴⁰ Marc J Gunter,⁶⁵ Robert W Haile,⁴ Jochen Hampe ,⁴¹ Michael Hoffmeister ,³¹ John L Hopper,^{42,66} Wan-Ling Hsu,¹⁵ Wen-Yi Huang ,²⁵ Thomas J Hudson,⁶⁷ Mazda Jenab ,⁶⁵ Mark A Jenkins,⁴² Amit D Joshi,^{51,53} Temitope O Keku,⁶⁸ Charles Kooperberg,¹ Tilman Kühn,⁵⁵ Sébastien Küry,³³ Loic Le Marchand,⁶⁴ Flavio Lejbkowitz,^{69,70,71} Christopher I Li,¹ Li Li,⁷² Wolfgang Lieb,⁷³ Annika Lindblom,^{74,75} Noralane M Lindor,⁷⁶ Satu Männistö,⁷⁷ Sanford D Markowitz,⁷⁸ Roger L Milne,^{42,59} Lorena Moreno,³⁰ Neil Murphy ,⁶⁵ Rami Nassir,⁷⁹ Kenneth Offit,^{80,81} Shuji Ogino,^{52,53,82,83} Salvatore Panico,⁸⁴ Patrick S Parfrey,⁸⁵ Rachel Pearlman,² Paul D P Pharoah,⁵⁸ Amanda I Phipps,^{1,86} Elizabeth A Platz,⁸⁷ John D Potter,¹ Ross L Prentice,¹ Lihong Qi,⁸⁸ Leon Raskin,⁸⁹ Gad Rennert,^{70,71,90} Hedy S Rennert,^{70,71,90} Elio Riboli,⁹¹ Clemens Schafmayer,⁹² Robert E Schoen,⁹³ Daniela Seminara,⁹⁴ Mingyang Song,^{49,51,95} Yu-Ru Su,¹ Catherine M Tangen,⁶¹ Stephen N Thibodeau,⁹⁶ Duncan C Thomas,⁶ Antonia Trichopoulou,^{97,98} Cornelia M Ulrich,³⁵ Kala Visvanathan,⁸⁷ Pavel Vodicka,^{99,100,101} Ludmila Vodickova,^{99,100,101} Veronika Vymetalkova,^{99,100,101} Korbinian Weigl,^{31,39,102} Stephanie J Weinstein,²⁵ Emily White,¹ Alicja Wolk,¹⁰³ Michael O Woods,¹⁰⁴ Anna H Wu,⁶ Goncalo R Abecasis,⁷ Deborah A Nickerson,¹⁰⁵ Peter C Scacheri,¹⁰⁶ Anshul Kundaje,^{32,107} Graham Casey,¹⁰⁸ Stephen B Gruber,^{109,110} Li Hsu,^{1,15} Victor Moreno,^{26,27,28} Richard B Hayes,¹¹¹ Polly A Newcomb,^{1,86} Ulrike Peters ,^{1,86}

► Additional material is published online only. To view please visit the journal online (<http://dx.doi.org/10.1136/gutjnl-2020-321534>).

For numbered affiliations see end of article.

Correspondence to

Dr Ulrike Peters, Public Health Sciences Division, Fred Hutchinson Cancer Research Center, Seattle, WA 98109, USA; upeters@fhcr.org

Albert de la Chapelle is deceased.

Received 23 April 2020

Revised 26 November 2020

Accepted 18 December 2020

Published Online First

25 February 2021



© 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.

To cite: Huyghe JR, Harrison TA, Bien SA, et al. *Gut* 2021;**70**:1325–1334.

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

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 ($p < 5 \times 10^{-8}$) 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 tumor 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.

Results We identified 13 loci that reached genome-wide significance ($p < 5 \times 10^{-8}$) and that were not reported by previous GWASs for overall CRC risk. Multiple lines of evidence support candidate genes at several of these loci. We detected substantial heterogeneity between anatomical subsites. Just over half (61) of 109 known and new risk variants showed no evidence for heterogeneity. In contrast, 22 variants showed association with distal CRC (including rectal cancer), but no evidence for association or an attenuated association with proximal CRC. For two loci, there was strong evidence for effects confined to proximal colon cancer.

Conclusion Genetic architectures of proximal and distal CRC are partly distinct. Studies of risk factors and mechanisms of carcinogenesis, and precision prevention strategies should take into consideration the anatomical subsite of the tumour.

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.^{2–5} 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 immunology. 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.^{7–10} Three decades ago, based on observed similarities between Lynch syndrome and proximal CRC, and between familial adenomatous polyposis and distal CRC, Buflin 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.

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.^{7 8 11 12} All participants provided written informed consent, and each study was approved by the relevant research ethics committee or institutional review board.

Colorectal 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 afore-mentioned 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.¹³ In brief, we phased all genotyping array data sets using SHAPEIT2¹⁴ and used the Michigan Imputation Server¹⁵ for imputation. Within each dataset, variants with an imputation accuracy $r^2 \geq 0.3$ and minor allele count ≥ 50 were tested for association with CRC case subgroup. 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.⁸ 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.¹⁶ 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 subsites, to a model in which ORs were constrained to be the same across tumour subsites. Third, inspired by reference,¹⁷ we used a multinomial logistic regression-based model selection approach to assess which configuration of tumour subsites 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 subsites (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 $\Delta BIC_i = BIC_i - BIC_{\min}$, where $BIC_{\min}$ is the BIC value for the best model. Models with $\Delta BIC_i \leq 2$ were considered to have substantial support and indistinguishable from the best model.¹⁸ For these variants, we do not report a single best model. Analyses were carried out using the VGAM R package.¹⁹ The list of index variants for previously published CRC risk signals is based on Huyghe *et al.*⁸

Pathway enrichment analyses

We used the Pascal programme to compute pathway enrichment score p values from genome-wide summary statistics.²⁰ The gene set library used comprises the combined KEGG,²¹ REACTOME²² and BIOCARTA²³ 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 subgroups 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 subsites 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 subsites, we identified 13 loci that mapped outside regions previously implicated by GWASs 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, $p = 3.8 \times 10^{-18}$) 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 subsites (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, $p = 4.9 \times 10^{-9}$), near *KLF14* on 7q32.3 (rs73161913, $p = 1.3 \times 10^{-9}$), 10q23.31 (rs7071258, $p = 8.4 \times 10^{-9}$), 19p13.3 (rs62131228, $p = 2.4 \times 10^{-8}$) and near *BMP7* on 20q13.31 (rs6014965, $p = 4.5 \times 10^{-9}$). The rectal cancer analysis identified an additional locus near *PYGL* on 14q22.1 (rs2861105, $p = 4.7 \times 10^{-9}$) 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_{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 (build 37)	Alleles (risk/other)	RAF (%)	OR	95% CI	P value	r ²	r ²	P _{het}	N cases	N controls
Colon	1p31.1	<i>PTGER3</i>	rs3124454	1	71 040 166	G/T	58.1	1.07	1.04 to 1.09	1.4E-08	0.926	6.1	0.38	32 002	64 159
Left-sided	2q21.3	<i>LCT</i>	rs1446585	2	136 407 479	G/A	39.9	1.07	1.04 to 1.10	3.3E-08	1.121	43.7	0.11	30 588	64 159
Proximal colon	3p22.2	<i>MLH1</i>	rs1800734†	3	37 034 946	A/G	24.7	1.15	1.11 to 1.19	3.8E-18	1.008	43.8	0.11	15 706	64 159
Colon	3p21.2	<i>STAB1</i> ; <i>TLR9</i> ; <i>NSCH</i>	rs353548	3	52 269 491	G/A	95.3	1.15	1.10 to 1.21	1.3E-08	0.975	0	0.48	32 002	64 159
Left-sided	5q32	<i>CDX1</i>	rs2302274†	5	149 546 426	G/A	47.8	1.07	1.04 to 1.09	4.9E-09	1.008	3.8	0.39	30 588	64 159
Left-sided	7q32.3	<i>KLF14</i> ; <i>LINC00513</i>	rs73161913†	7	130 607 779	G/A	94.3	1.16	1.10 to 1.22	1.3E-09	0.975	0	0.79	30 588	64 159
Left-sided	10q23.31	<i>PANK1</i> ; <i>KIF20B</i>	rs7071258†	10	91 574 624	A/G	21.6	1.08	1.05 to 1.11	8.4E-09	0.993	0	0.71	30 588	64 159
Rectal	14q22.1	<i>PYGL</i> ; <i>NIN</i> ; <i>ABHD12B</i>	rs2861105†	14	51 359 658	G/T	21.5	1.11	1.07 to 1.15	4.7E-09	0.983	50.5	0.07	16 212	64 159
Proximal colon	14q32.12		rs61975764	14	93 014 929	G/A	55.3	1.08	1.05 to 1.11	2.8E-08	0.987	0	0.71	15 706	64 159
Proximal colon	14q32.2	<i>BCL11B</i>	rs80158569†	14	99 782 937	A/G	7.5	1.18	1.12 to 1.24	8.6E-11	0.899	29.9	0.21	15 706	64 159
Left-sided	19p13.3	<i>STK11</i> ; <i>SBNO2</i>	rs62131228	19	1 157 642	G/A	98.1	1.28	1.17 to 1.40	2.4E-08	0.788	0	0.95	29 632	63 385
Left-sided	20q13.31	<i>BMP7</i>	rs6014965†	20	55 831 203	A/G	55.4	1.07	1.04 to 1.09	4.5E-09	0.995	10.5	0.35	30 588	64 159
Colon	22q13.31	<i>FAM118A</i> ; <i>FBLN1</i>	rs736037	22	45 724 999	A/G	28.6	1.07	1.04 to 1.09	2.8E-08	1.015	0	0.74	32 002	64 159

Lead variant is the most significant variant at the locus. Reference single nucleotide polymorphism (SNP) cluster ID (rsID) based on NCB dbSNP Build 152. Alleles are on the + strand. All p values reported in this table are from a sample size-weighted fixed-effects meta-analysis of logistic regression-based likelihood-ratio test results. Reported imputation qualities r^2 are effective sample size (N_{eff})-weighted means across the six data sets, where $N_{\text{eff}} = 4 \times (1/N_{\text{het}} + 1/N_{\text{hom}})$. The r^2 statistic measures heterogeneity on a scale of 0–100%. P_{het} is the p value from Cochran's Q test for heterogeneity.

*Colon: proximal–distal colon; colon: unspecified site; left-sided: distal colon–rectal. Details of tumour site definitions including ICD-9 codes are given in the Methods section and online supplemental materials.

†Variant attained Bonferroni-adjusted genome-wide significance ($5E-08/5 = 1E-08$), corrected for the number of CRC case subgroups analysed.

Chr., chromosome; CRC, colorectal cancer; RAF, risk allele frequency.

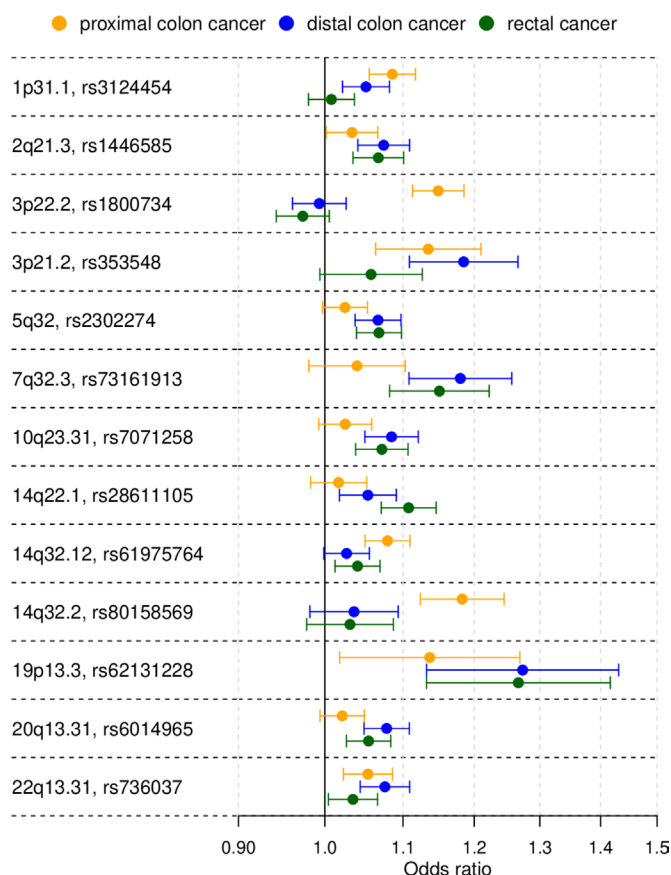


Figure 1 Primary tumour site-specific associations for the lead single nucleotide polymorphisms (SNPs) of the 13 colorectal cancer risk loci not reported in previous genome-wide association studies. The forest plot shows the (log-additive) OR estimates together with 95% CIs. For clarity, this figure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

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 $p_{\text{het}} < 0.05$, and after Holm-Bonferroni correction for multiple testing, 14 variants showing strong evidence of heterogeneity remained significant ($p_{\text{het}} < 4.6 \times 10^{-4}$). These included 10 variants previously reported in GWASs for overall CRC risk.

For 17 out of the 48 variants with $p_{\text{het}} < 0.05$, 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 ($p_{\text{het}} < 4.6 \times 10^{-4}$), including the following previously reported loci: *C11orf53-COLCA1-COLCA2* on 11q23.1 ($p_{\text{het}} = 6.0 \times 10^{-14}$), *APC* on 5q22.2 ($p_{\text{het}} = 2.3 \times 10^{-10}$), *GATA3* on 10p14 ($p_{\text{het}} = 1.7 \times 10^{-8}$), *CTNNB1* on 3p22.1 ($p_{\text{het}} = 9.8 \times 10^{-8}$), *RAB40B-METRLN* on 17q25.3 ($p_{\text{het}} = 3.6 \times 10^{-6}$) and *CDKN1A* on 6p21.2 ($p_{\text{het}} = 1.6 \times 10^{-4}$). 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, *VITI1A* on 10q25.2, two independent signals near *POLD3* on 11q13.4, and *BMP4* on 14q22.2.

For 5 out of the 49 variants with $p_{\text{het}} < 0.05$, a model with association with colon cancer risk, but no association with rectal

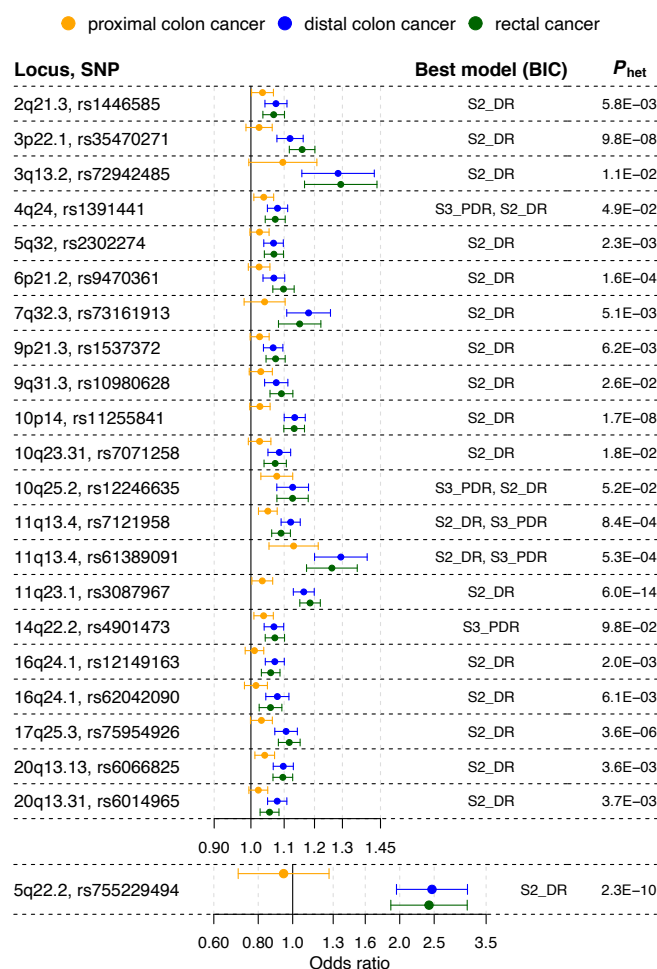


Figure 2 Loci showing association with risk of distal colorectal cancer (ie, distal colon+rectal), but attenuated or no evidence for association with proximal colon cancer risk. The forest plot shows the (log-additive) OR estimates for the lead single nucleotide polymorphisms (SNPs) at the loci, together 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 subsites.

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/B-NFKBIL1-TNF* on 6p21.33, *NOS1* on 12q24.22 and *LINC00673* 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 3p22.2 ($p_{\text{het}} = 5.4 \times 10^{-19}$), and *BCL11B* ($p_{\text{het}} = 1.5 \times 10^{-5}$) 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

loci *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 subsite, reached at least suggestive significance levels for other subsites (online supplemental table 8). Several of the Bonferroni-significant pathways related to transforming growth factor β (TGF β) 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 adequate statistical power to discover new loci and variants with risk effects limited to tumours for certain anatomical subsites, and to compare allelic effect sizes across anatomical subsites.

Our CRC case subgroup meta-analyses identified 13 additional genome-wide significant CRC risk loci that, due to substantial allelic effect heterogeneity between anatomical subsites, were not detected in larger, previously published GWASs for overall CRC risk.^{8,9} In fact, the only way to discover certain loci and risk variants with case subgroup-specific allelic effects is via analysis of homogeneous case subgroups.²⁴ For example, p values for rs1800734 and rs80158569 were ~ 18 and ~ 5 powers of 10, respectively, more significant in the proximal colon analysis compared with in our overall CRC analysis. While follow-up studies are needed to uncover the causal variant(s), biological mechanism and target gene, multiple lines of evidence support strong candidate target genes at many of the new loci, including genes *MLH1*, *BCL11B*, *RIN3*, *CDX1*, *LCT*, *KLF14*, *BMP7*, *PYGL* and *PTGER3*.

At the *MLH1* gene promoter region on 3p22.2, associated to proximal colon cancer, previous studies have reported strong and robust associations between the common single nucleotide polymorphism (SNP) rs1800734, and CRC with high microsatellite instability (MSI-H).^{25,26} Rare deleterious nonsynonymous germline mutations in the DNA mismatch repair (MMR) gene *MLH1* are a frequent cause of Lynch syndrome (OMIM #609310). The risk allele of the likely causal SNP rs1800734 is strongly associated with *MLH1* promoter hypermethylation and loss of *MLH1* protein in CRC tumours.²⁶ The mechanisms of *MLH1* promoter hypermethylation and subsequent gene silencing may account for most CRC tumours with defective DNA MMR and MSI-H.²⁷

At the highly localised, proximal colon-specific association signal on 14q32.2, lead SNP rs80158569 is located in a colonic crypt enhancer and overlaps with multiple transcription factor binding sites, making it a strong candidate causal variant. Nearby gene *BCL11B* encodes a transcription factor that is required for normal T cell development,^{28,29} and that is a SWI/SNF complex subunit.³⁰ *BCL11B* acts as a haploinsufficient tumour suppressor in T-cell acute lymphoblastic leukaemia.^{31,32} Experimental work suggests that impairment of *Bcl11b* promotes intestinal

tumourigenesis in mice and humans through deregulation of the Wnt/ β -catenin pathway.³³

At locus 14q32.12, lead SNP rs61975764 showed the strongest association evidence in the proximal colon analysis and attenuated effects for other tumour locations. Genotype-Tissue Expression (GTEx) data show that rs61975764 is an eQTL for gene *Ras* and *Rab* interactor 3 (*RIN3*) in transverse colon tissue. *RIN3* functions as a *RAB5* and *RAB31* guanine nucleotide exchange factor involved in endocytosis.^{34,35}

At locus 5q32, 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.³⁶ *CDX1* has been shown to inhibit human colon cancer cell proliferation by blocking β -catenin/T-cell factor transcriptional activity.³⁷

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 $r^2=0.854$) 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.³⁸ The protective effect conferred by the lactase persistence genotype is likely mediated by dairy products and calcium which are known protective factors for CRC.³⁹ 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^{-9} , 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.⁴⁰ 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 20q13.31 are involved in TGF β 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*.⁸ The imprinted gene *KLF14* shows monoallelic maternal expression, and is induced by TGF β to transcriptionally corepress the TGF β receptor 2 (*TGFBR2*) gene.⁴¹ A *cis*-eQTL for *KLF14*, uncorrelated with our lead SNP rs73161913, acts as a master regulator related to multiple metabolic phenotypes,^{42,43} and a nearby independent variant is associated to basal cell carcinoma.⁴⁴ 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.⁴⁵ In CRC tumours,

BMP7 expression correlates with parameters of pathological aggressiveness such as liver metastasis and poor prognosis.⁴⁶

On 14q22.1, the single locus identified only in the rectal cancer analysis, GTEx data show that, in gastrointestinal tissues, lead SNP rs28611105 localises with a *cis*-eQTL coregulating expression of genes *PYGL*, *ABHD12B* and *NIN*. We reported an association between genetically predicted glycogen phosphorylase L (*PYGL*) expression and CRC risk in a transcriptome-wide association study.⁴⁷ This glycogen 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 (PGE2), 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 PGE2 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 sublocations, 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 TGF β 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 sublocations. 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 MAP5 (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 of Research, Kaiser Permanente Northern California, Oakland, California, USA

²⁰Department 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'Hospitalet de Llobregat, Barcelona, Spain

- ²⁵Division of Cancer Epidemiology and Genetics, National Cancer Institute, Bethesda, Maryland, USA
- ²⁶Cancer Prevention and Control Program, Catalan Institute of Oncology - IDIBELL, L'Hospitalet de Llobregat, Barcelona, Spain
- ²⁷CIBER Epidemiología y Salud Pública (CIBERESP), Madrid, Spain
- ²⁸Department of Clinical Sciences, Faculty of Medicine, University of Barcelona, Barcelona, Spain
- ²⁹Division of Epidemiology and Community Health, University of Minnesota, Minneapolis, Minnesota, USA
- ³⁰Gastroenterology Department, Hospital Clínic, Institut d'Investigacions Biomèdiques August Pi i Sunyer (IDIBAPS), Centro de Investigación Biomédica en Red de Enfermedades Hepáticas y Digestivas (CIBEREHD), University of Barcelona, Barcelona, Spain
- ³¹Division of Clinical Epidemiology and Aging Research, German Cancer Research Centre (DKFZ), Heidelberg, Germany
- ³²Department of Genetics, Stanford University, Stanford, California, USA
- ³³Service de Génétique Médicale, Centre Hospitalier Universitaire (CHU) de Nantes, Nantes, France
- ³⁴Leeds Institute of Medical Research at St James's, University of Leeds, Leeds, UK
- ³⁵Huntsman Cancer Institute and Department of Population Health Sciences, University of Utah Health, Salt Lake City, Utah, USA
- ³⁶Department of Epidemiology, German Institute of Human Nutrition (DIfE), Potsdam-Rehbrücke, Germany
- ³⁷Institut Gustave Roussy, Université Paris-Saclay, Villejuif, France
- ³⁸Division of Preventive Oncology, German Cancer Research Centre (DKFZ) and National Center for Tumor Diseases (NCT), Heidelberg, Germany
- ³⁹German Cancer Consortium (DKTK), German Cancer Research Center (DKFZ), Heidelberg, Germany
- ⁴⁰Institute of Cancer Research, Department of Medicine I, Medical University of Vienna, Vienna, Austria
- ⁴¹Department of Medicine I, University Hospital Dresden, Technische Universität Dresden (TU Dresden), Dresden, Germany
- ⁴²Centre for Epidemiology and Biostatistics, Melbourne School of Population and Global Health, The University of Melbourne, Melbourne, Victoria, Australia
- ⁴³Colorectal Oncogenomics Group, Genetic Epidemiology Laboratory, Department of Clinical Pathology, The University of Melbourne, Melbourne, Victoria, Australia
- ⁴⁴Genomic Medicine and Family Cancer Clinic, Royal Melbourne Hospital, Melbourne, Victoria, Australia
- ⁴⁵Institute for Health Research, Kaiser Permanente Colorado, Denver, Colorado, USA
- ⁴⁶Division of Research, Kaiser Permanente Medical Care Program, Oakland, California, USA
- ⁴⁷Behavioral and Epidemiology Research Group, American Cancer Society, Atlanta, Georgia, USA
- ⁴⁸Division of Clinical Epidemiology, German Cancer Research Centre (DKFZ), Heidelberg, Germany
- ⁴⁹Division of Gastroenterology, Massachusetts General Hospital and Harvard Medical School, Boston, Massachusetts, USA
- ⁵⁰Channing Division of Network Medicine, Brigham and Women's Hospital and Harvard Medical School, Boston, Massachusetts, USA
- ⁵¹Clinical and Translational Epidemiology Unit, Massachusetts General Hospital and Harvard Medical School, Boston, Massachusetts, USA
- ⁵²Broad Institute of Harvard and MIT, Cambridge, Massachusetts, USA
- ⁵³Department of Epidemiology, Harvard T.H. Chan School of Public Health, Harvard University, Boston, Massachusetts, USA
- ⁵⁴Department of Immunology and Infectious Diseases, Harvard T.H. Chan School of Public Health, Harvard University, Boston, Massachusetts, USA
- ⁵⁵Division of Cancer Epidemiology, German Cancer Research Centre (DKFZ), Heidelberg, Germany
- ⁵⁶Cancer Epidemiology Group, University Medical Centre Hamburg-Eppendorf, University Cancer Centre Hamburg (UCCH), Hamburg, Germany
- ⁵⁷Department of Cancer Biology and Genetics and the Comprehensive Cancer Center, The Ohio State University, Columbus, Ohio, USA
- ⁵⁸Department of Public Health and Primary Care, University of Cambridge, Cambridge, UK
- ⁵⁹Cancer Epidemiology Division, Cancer Council Victoria, Melbourne, Victoria, Australia
- ⁶⁰Lunenfeld-Tanenbaum Research Institute, Mount Sinai Hospital, University of Toronto, Toronto, Ontario, Canada
- ⁶¹SWOG Statistical Center, Fred Hutchinson Cancer Research Center, Seattle, Washington, USA
- ⁶²Clinical Research Division, Fred Hutchinson Cancer Research Center, Seattle, Washington, USA
- ⁶³Department of Medicine, University of Washington School of Medicine, Seattle, Washington, USA
- ⁶⁴University of Hawai'i Cancer Center, Honolulu, Hawaii, USA
- ⁶⁵Nutrition and Metabolism Section, International Agency for Research on Cancer, World Health Organization, Lyon, France
- ⁶⁶Department of Epidemiology, School of Public Health and Institute of Health and Environment, Seoul National University, Seoul, South Korea
- ⁶⁷Ontario Institute for Cancer Research, Toronto, Ontario, Canada
- ⁶⁸Center for Gastrointestinal Biology and Disease, University of North Carolina, Chapel Hill, North Carolina, USA
- ⁶⁹The Clalit Health Services, Personalized Genomic Service, Carmel Medical Center, Haifa, Israel
- ⁷⁰Department of Community Medicine and Epidemiology, Lady Davis Carmel Medical Center, Haifa, Israel
- ⁷¹Clalit National Cancer Control Center, Haifa, Israel
- ⁷²Department of Family Medicine, University of Virginia, Charlottesville, Virginia, USA
- ⁷³Institute of Epidemiology, PopGen Biobank, Christian-Albrechts-University Kiel, Kiel, Germany
- ⁷⁴Department of Clinical Genetics, Karolinska University Hospital, Stockholm, Sweden
- ⁷⁵Department of Molecular Medicine and Surgery, Karolinska Institutet, Stockholm, Sweden
- ⁷⁶Department of Health Science Research, Mayo Clinic, Scottsdale, Arizona, USA
- ⁷⁷Department of Public Health Solutions, National Institute for Health and Welfare, Helsinki, Finland
- ⁷⁸Departments of Medicine and Genetics, Case Comprehensive Cancer Center, Case Western Reserve University and University Hospitals of Cleveland, Cleveland, Ohio, USA
- ⁷⁹Department of Pathology, School of Medicine, Umm Al-Qura'a University, Mecca, Saudi Arabia
- ⁸⁰Clinical Genetics Service, Department of Medicine, Memorial Sloan Kettering Cancer Center, New York, New York, USA
- ⁸¹Department of Medicine, Weill Cornell Medical College, New York, New York, USA
- ⁸²Department of Oncologic Pathology, Dana-Farber Cancer Institute, Boston, Massachusetts, USA
- ⁸³Program in MPE Molecular Pathological Epidemiology, Department of Pathology, Brigham and Women's Hospital and Harvard Medical School, Boston, Massachusetts, USA
- ⁸⁴Dipartimento di Medicina Clinica e Chirurgia, University of Naples Federico II, Naples, Italy
- ⁸⁵Clinical Epidemiology Unit, Faculty of Medicine, Memorial University of Newfoundland, St. John's, Newfoundland, Canada
- ⁸⁶Department of Epidemiology, University of Washington, Seattle, Washington, USA
- ⁸⁷Department of Epidemiology, Johns Hopkins Bloomberg School of Public Health, Johns Hopkins University, Baltimore, Maryland, USA
- ⁸⁸Department of Public Health Sciences, School of Medicine, University of California Davis, Davis, California, USA
- ⁸⁹Division of Epidemiology, Vanderbilt Epidemiology Center, Vanderbilt University School of Medicine, Nashville, Tennessee, USA
- ⁹⁰Ruth and Bruce Rappaport Faculty of Medicine, Technion-Israel Institute of Technology, Haifa, Israel
- ⁹¹School of Public Health, Imperial College London, London, UK
- ⁹²Department of General Surgery, University Hospital Rostock, Rostock, Germany
- ⁹³Department of Medicine and Epidemiology, University of Pittsburgh Medical Center, Pittsburgh, Pennsylvania, USA
- ⁹⁴Division of Cancer Control and Population Sciences, National Cancer Institute, Bethesda, Maryland, USA
- ⁹⁵Department of Nutrition, Harvard T.H. Chan School of Public Health, Harvard University, Boston, Massachusetts, USA
- ⁹⁶Division of Laboratory Genetics, Department of Laboratory Medicine and Pathology, Mayo Clinic, Rochester, Minnesota, USA
- ⁹⁷Hellenic Health Foundation, Athens, Greece
- ⁹⁸WHO Collaborating Center for Nutrition and Health, Unit of Nutritional Epidemiology and Nutrition in Public Health, Department of Hygiene, Epidemiology and Medical Statistics, School of Medicine, National and Kapodistrian University of Athens, Athens, Greece
- ⁹⁹Department of Molecular Biology of Cancer, Institute of Experimental Medicine of the Czech Academy of Sciences, Prague, Czech Republic
- ¹⁰⁰Institute of Biology and Medical Genetics, First Faculty of Medicine, Charles University, Prague, Czech Republic
- ¹⁰¹Faculty of Medicine and Biomedical Center in Pilsen, Charles University, Pilsen, Czech Republic
- ¹⁰²Medical Faculty, University of Heidelberg, Heidelberg, Germany
- ¹⁰³Institute of Environmental Medicine, Karolinska Institutet, Stockholm, Sweden
- ¹⁰⁴Discipline of Genetics, Memorial University of Newfoundland, St. John's, Newfoundland, Canada
- ¹⁰⁵Department of Genome Sciences, University of Washington, Seattle, Washington, USA
- ¹⁰⁶Department of Genetics and Genome Sciences, Case Western Reserve University School of Medicine, Case Comprehensive Cancer Center, Cleveland, Ohio, USA
- ¹⁰⁷Department of Computer Science, Stanford University, Stanford, California, USA
- ¹⁰⁸Center for Public Health Genomics, University of Virginia, Charlottesville, Virginia, USA

¹⁰⁹Department of Preventive Medicine, USC Norris Comprehensive Cancer Center, University of Southern California Keck School of Medicine, Los Angeles, California, USA

¹¹⁰City of Hope National Medical Center, Duarte, California, USA

¹¹¹Division of Epidemiology, Department of Population Health, New York University School of Medicine, New York, New York, USA

Twitter Daniel D Buchanan @dan_buchanan, Sergi Castellvi-Bel @scastellvibel and Mazda Jenab @mazda_j

Contributors JRH, TAH, SAB, HH, JCF, SLS, DVC, JAB, AJC, BD, DD, SH, LI, VP, AP-C, LCS, FRS, MLS, AET, FJBvD, BVG, AA, DA, MHA, KA, CA-C, VA, SIB, SB, DTB, JB, HBoeing, M-CB-R, HBrenner, SBrezina, SBuch, DDB, AB-H, BJC, PTC, PC, AC, SC-B, ATC, JC-C, SJC, AdIC, DFE, DRE, EJMf, MG, SJG, WJG, GGG, PJG, WMG, JSG, AG, MJG, RWH, JH, MH, JLH, W-YH, TJH, MJ, MAJ, ADJ, TOK, CK, TK, SK, LLM, FL, CIL, LL, WL, AL, NML, SM, SDM, RLM, LM, NM, RN, KO, SO, SP, PSP, RP, PDPP, AIP, EAP, JDP, RLP, LQ, LR, GR, HSR, ER, CS, RES, DS, MS, CMT, SNT, DCT, AT, CMU, KV, PV, LV, VV, KW, SJW, EW, AW, MOW, AHW, GRA, DAN, PCS, AK, GC, SBG, LH, VM, RBH, PAN and UP conceived and designed the study. JRH, TAH, SAB, SLS, DVC, SC, CQ, YL, RB, HMK, DML, FRS, BB, KRC, W-LH, Y-RS, AK, LH and UP analysed the data. JRH, TAH, HH, JCF, JAB, AJC, BD, SH, LI, HMK, VP, AP-C, LCS, MLS, AET, FJBvD, BVG, AA, DA, MHA, KA, CA-C, VA, MCB, SIB, SB, DTB, JB, HBoeing, M-CB-R, HBrenner, SBrezina, SBuch, DDB, AB-H, BJC, PTC, PC, AC, SC-B, ATC, JC-C, SJC, AdIC, DFE, DRE, EJMf, MG, SJG, WJG, GGG, PJG, WMG, JSG, AG, MJG, RWH, JH, MH, JLH, W-LH, W-YH, TJH, MJ, MAJ, ADJ, TOK, CK, TK, SK, LLM, FL, CIL, LL, WL, AL, NML, SM, SDM, RLM, LM, NM, RN, KO, SO, SP, PSP, RP, PDPP, AIP, EAP, JDP, RLP, LQ, LR, GR, HSR, ER, CS, RES, MS, Y-RS, CMT, SNT, DCT, AT, CMU, KV, PV, LV, VM, KW, SJW, EW, AW, MOW, AHW, GRA, DAN, PCS, AK, GC, SBG, VM, RBH, PAN and UP contributed reagents/materials/analysis tools. JRH, TH and UP wrote the first draft. All authors reviewed the manuscript for intellectual content and approved the final version of the manuscript. UP supervised the study.

Funding This work was supported by grants from the National Cancer Institute (NCI), National Institutes of Health (NIH), US Department of Health and Human Services (U01 CA164930, U01 CA137088, R01 CA059045, R21 CA191312, R01 CA201407, P30 CA015704). Genotyping services were provided by the Center for Inherited Disease Research (CIDR; X01-HG008596 and X01-HG007585). CIDR is fully funded through a federal contract from the NIH to the Johns Hopkins University, contract HHSN2682012000081. The full list of funding and acknowledgements can be found in the supplemental file.

Disclaimer Where authors are identified as personnel of the International Agency for Research on Cancer/WHO, the authors alone are responsible for the views expressed in this article and they do not necessarily represent the decisions, policy or views of the International Agency for Research on Cancer/WHO.

Competing interests None declared.

Patient consent for publication Not required.

Provenance and peer review Not commissioned; externally peer reviewed.

Data availability statement Data are available in a public controlled access repository. All genotype data analyzed in this study have been previously published and have been deposited in the database of Genotypes and Phenotypes (dbGaP), which is hosted by the National Center for Biotechnology Information (NCBI) of the US National Institutes of Health (NIH), under accession numbers phs001415.v1.p1, phs001315.v1.p1, phs001078.v1.p1, and phs001903.v1.p1. The UK Biobank resource was accessed through application number 8614. Bioinformatic analyses included public, open access colorectal epigenomic data that were retrieved from the NCBI Gene Expression Omnibus (GEO) database under accession numbers GSE77737 and GSE36401. For all above datasets embargo release dates have passed.

Supplemental material This content has been supplied by the author(s). It has not been vetted by BMJ Publishing Group Limited (BMJ) and may not have been peer-reviewed. Any opinions or recommendations discussed are solely those of the author(s) and are not endorsed by BMJ. BMJ disclaims all liability and responsibility arising from any reliance placed on the content. Where the content includes any translated material, BMJ does not warrant the accuracy and reliability of the translations (including but not limited to local regulations, clinical guidelines, terminology, drug names and drug dosages), and is not responsible for any error and/or omissions arising from translation and adaptation or otherwise.

Open access This is an open access article distributed in accordance with the Creative Commons Attribution Non Commercial (CC BY-NC 4.0) license, which permits others to distribute, remix, adapt, build upon this work non-commercially, and license their derivative works on different terms, provided the original work is properly cited, appropriate credit is given, any changes made indicated, and the use is non-commercial. See: <http://creativecommons.org/licenses/by-nc/4.0/>.

ORCID iDs

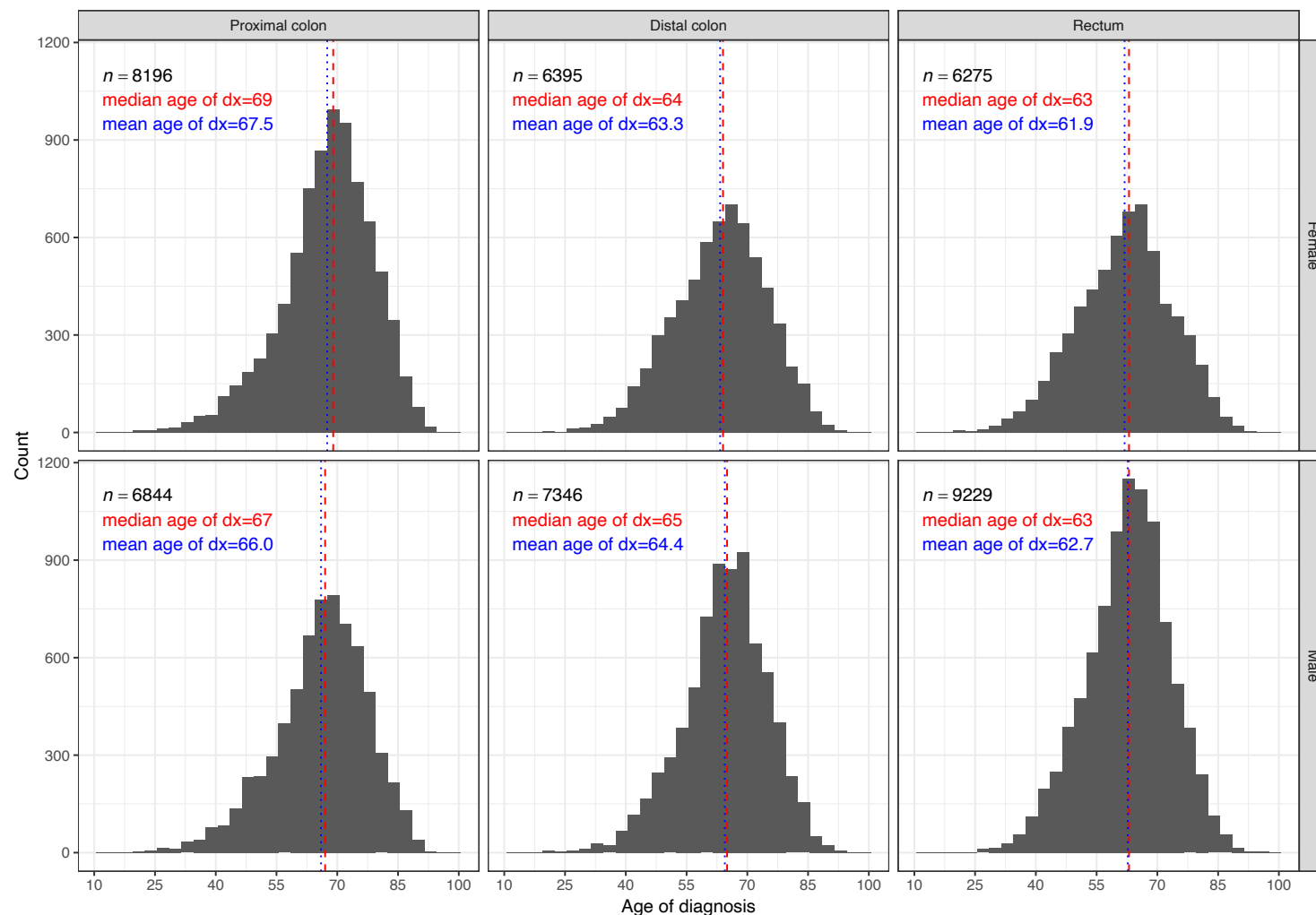
Jeroen R Huyghe <http://orcid.org/0000-0001-6027-9806>
Hermann Brenner <http://orcid.org/0000-0002-6129-1572>

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/0000-0002-4440-3368>
Mazda Jenab <http://orcid.org/0000-0002-0573-1852>
Neil Murphy <http://orcid.org/0000-0003-3347-8249>
Ulrike Peters <http://orcid.org/0000-0001-5666-9318>

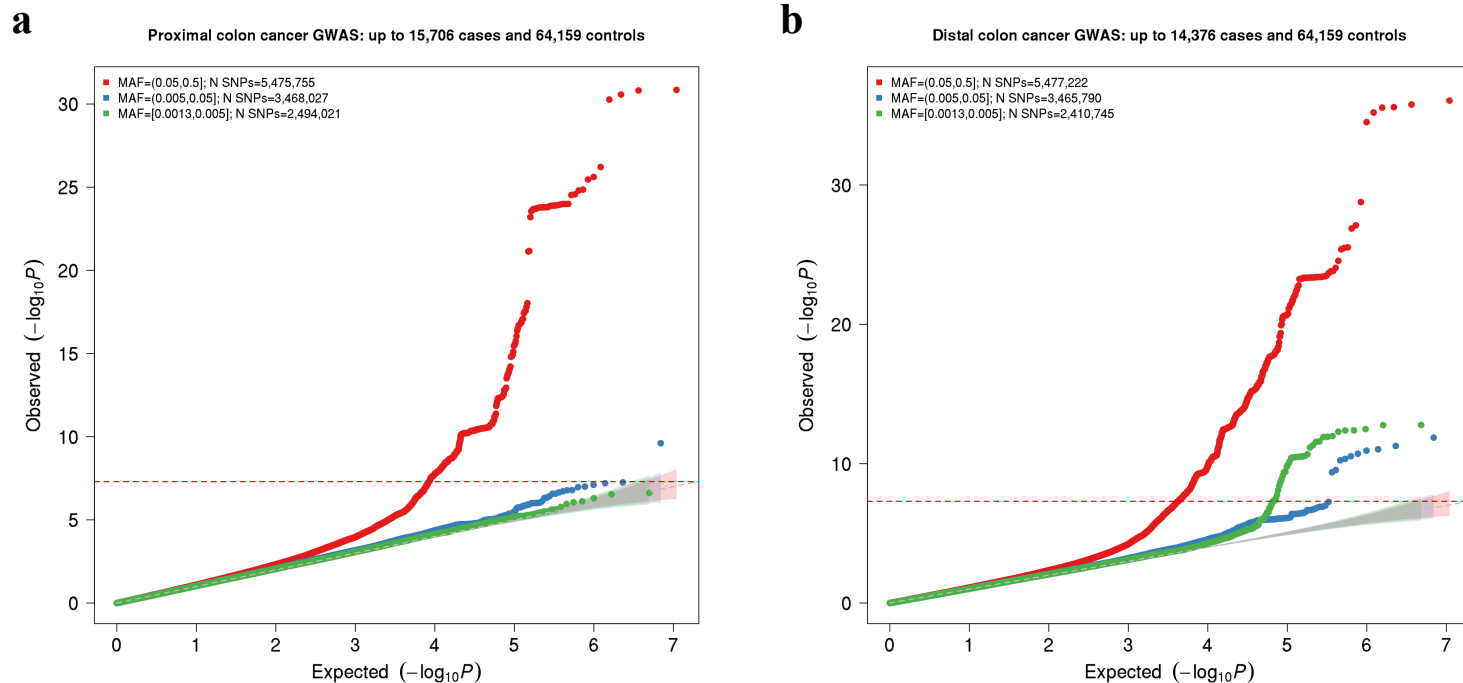
REFERENCES

- 1 American Cancer Society. Cancer statistics center. Available: <http://cancerstatisticscenter.cancer.org> [Accessed 21 Apr 2020].
- 2 Buffill JA. Colorectal cancer: evidence for distinct genetic categories based on proximal or distal tumor location. *Ann Intern Med* 1990;113:779–88.
- 3 Iacopetta B. Are there two sides to colorectal cancer? *Int J Cancer* 2002;101:403–8.
- 4 Carethers JM. One colon lumen but two organs. *Gastroenterology* 2011;141:411–2.
- 5 Yamauchi M, Lochhead P, Morikawa T, et al. Colorectal cancer: a tale of two sides or a continuum? *Gut* 2012;61:794–7.
- 6 Lichtenstein P, Holm NV, Verkasalo PK, et al. Environmental and heritable factors in the causation of cancer—analyses of cohorts of twins from Sweden, Denmark, and Finland. *N Engl J Med* 2000;343:78–85.
- 7 Schmit SL, Edlund CK, Schumacher FR, et al. Novel common genetic susceptibility loci for colorectal cancer. *J Natl Cancer Inst* 2019;111:146–57.
- 8 Huyghe JR, Bien SA, Harrison TA, et al. Discovery of common and rare genetic risk variants for colorectal cancer. *Nat Genet* 2019;51:76–87.
- 9 Law PJ, Timofeeva M, Fernandez-Rozadilla C, et al. Association analyses identify 31 new risk loci for colorectal cancer susceptibility. *Nat Commun* 2019;10:2154.
- 10 Lu Y, Kweon S-S, Tanikawa C, et al. Large-Scale genome-wide association study of East Asians identifies loci associated with risk for colorectal cancer. *Gastroenterology* 2019;156:1455–66.
- 11 Peters U, Jiao S, Schumacher FR, et al. Identification of genetic susceptibility loci for colorectal tumors in a genome-wide meta-analysis. *Gastroenterology* 2013;144:799–807.
- 12 Schumacher FR, Schmit SL, Jiao S, et al. Genome-Wide association study of colorectal cancer identifies six new susceptibility loci. *Nat Commun* 2015;6:7138.
- 13 McCarthy S, Das S, Kretschmar W, et al. A reference panel of 64,976 haplotypes for genotype imputation. *Nat Genet* 2016;48:1279–83.
- 14 Delaneau O, Howie B, Cox AJ, et al. Haplotype estimation using sequencing reads. *Am J Hum Genet* 2013;93:687–96.
- 15 Das S, Forer L, Schönerr S, et al. Next-Generation genotype imputation service and methods. *Nat Genet* 2016;48:1284–7.
- 16 Willer CJ, Li Y, Abecasis GR. METAL: fast and efficient meta-analysis of genomewide association scans. *Bioinformatics* 2010;26:2190–1.
- 17 Cross-Disorder Group of the Psychiatric Genomics Consortium. Identification of risk loci with shared effects on five major psychiatric disorders: a genome-wide analysis. *Lancet* 2013;381:1371–9.
- 18 Burnham KP, Anderson DR. Multimodel inference: understanding AIC and BIC in model selection. *Social Methods Res* 2004;33:261–304.
- 19 Yee TW. The VGAM Package for Categorical Data Analysis. *J Stat Softw* 2010;32.
- 20 Lamparter D, Marbach D, Rueedi R, et al. Fast and rigorous computation of gene and pathway scores from SNP-based summary statistics. *PLoS Comput Biol* 2016;12:e1004714.
- 21 Kanehisa M, Goto S. KEGG: Kyoto encyclopedia of genes and genomes. *Nucleic Acids Res* 2000;28:27–30.
- 22 Croft D, O’Kelly G, Wu G, et al. Reactome: a database of reactions, pathways and biological processes. *Nucleic Acids Res* 2011;39:D691–7.
- 23 Nishimura D. BioCarta. *Biotech Software & Internet Report* 2001;2:117–20.
- 24 T aylor M, Markus H, Lewis CM. Homogeneous case subgroups increase power in genetic association studies. *Eur J Hum Genet* 2015;23:863–9.
- 25 Raptis S, Mrkonjic M, Green RC, et al. MLH1 -93G>A promoter polymorphism and the risk of microsatellite-unstable colorectal cancer. *J Natl Cancer Inst* 2007;99:463–74.
- 26 Mrkonjic M, Roslin NM, Greenwood CM, et al. Specific variants in the MLH1 gene region may drive DNA methylation, loss of protein expression, and MSI-H colorectal cancer. *PLoS One* 2010;5:e13314.
- 27 Cunningham JM, Christensen ER, Tester DJ, et al. Hypermethylation of the hMLH1 promoter in colon cancer with microsatellite instability. *Cancer Res* 1998;58:3455–60.
- 28 Avram D, Califano D. The multifaceted roles of Bcl11b in thymic and peripheral T cells: impact on immune diseases. *J Immunol* 2014;193:2059–65.
- 29 Punwani D, Zhang Y, Yu J, et al. Multisystem anomalies in severe combined immunodeficiency with mutant Bcl11b. *N Engl J Med* 2016;375:2165–76.
- 30 Kadoch C, Hargreaves DC, Hodges C, et al. Proteomic and bioinformatic analysis of mammalian SWI/SNF complexes identifies extensive roles in human malignancy. *Nat Genet* 2013;45:592–601.

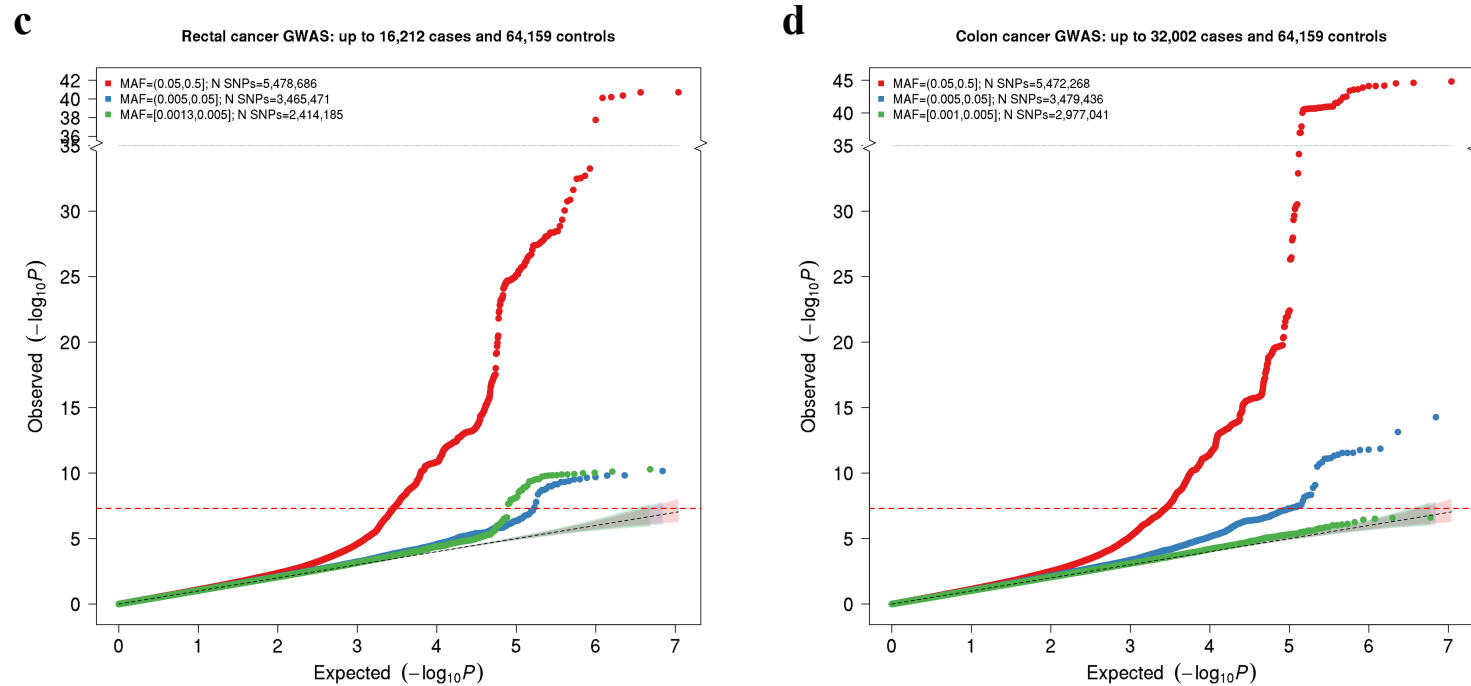
- 31 Gutierrez A, Kentsis A, Sanda T, *et al.* The BCL11B tumor suppressor is mutated across the major molecular subtypes of T-cell acute lymphoblastic leukemia. *Blood* 2011;118:4169–73.
- 32 Neumann M, Vosberg S, Schlee C, *et al.* Mutational spectrum of adult T-ALL. *Oncotarget* 2015;6:2754–66.
- 33 Sakamaki A, Katsuragi Y, Otsuka K, *et al.* Bcl11b SWI/SNF-complex subunit modulates intestinal adenoma and regeneration after γ -irradiation through Wnt/ β -catenin pathway. *Carcinogenesis* 2015;36:622–31.
- 34 Kajihito H, Saito K, Tsujita K, *et al.* RIN3: a novel Rab5 GEF interacting with amphiphysin II involved in the early endocytic pathway. *J Cell Sci* 2003;116:4159–68.
- 35 Kajihito H, Sakurai K, Minoda T, *et al.* Characterization of RIN3 as a guanine nucleotide exchange factor for the Rab5 subfamily GTPase Rab31. *J Biol Chem* 2011;286:24364–73.
- 36 Jones MF, Hara T, Francis P, *et al.* The CDX1-microRNA-215 axis regulates colorectal cancer stem cell differentiation. *Proc Natl Acad Sci U S A* 2015;112:E1550–8.
- 37 Guo R-J, Huang E, Ezaki T, *et al.* Cdx1 inhibits human colon cancer cell proliferation by reducing β -catenin/T-cell factor transcriptional activity. *J Biol Chem* 2004;279:36865–75.
- 38 Räsänen H, Forsblom C, Enattah NS, *et al.* The C/C-13910 genotype of adult-type hypolactasia is associated with an increased risk of colorectal cancer in the Finnish population. *Gut* 2005;54:643–7.
- 39 World Cancer Research Fund/American Institute for Cancer Research. Continuous update project expert report 2018. diet, nutrition, physical activity and colorectal cancer. Available: dietandcancerreport.org
- 40 Campbell CD, Ogburn EL, Lunetta KL, *et al.* Demonstrating stratification in a European American population. *Nat Genet* 2005;37:868–72.
- 41 Truty MJ, Lomberg G, Fernandez-Zapico ME, *et al.* Silencing of the transforming growth factor- β (TGF β) receptor II by Kruppel-like factor 14 underscores the importance of a negative feedback mechanism in TGF β signaling. *J Biol Chem* 2009;284:6291–300.
- 42 Small KS, Hedman AK, Grundberg E, *et al.* Identification of an imprinted master trans regulator at the Klf14 locus related to multiple metabolic phenotypes. *Nat Genet* 2011;43:1040–4.
- 43 Small KS, Todorčević M, Civelek M, *et al.* Regulatory variants at Klf14 influence type 2 diabetes risk via a female-specific effect on adipocyte size and body composition. *Nat Genet* 2018;50:572–80.
- 44 Stacey SN, Sulem P, Masson G, *et al.* New common variants affecting susceptibility to basal cell carcinoma. *Nat Genet* 2009;41:909–14.
- 45 Eikesdal HP, Becker LM, Teng Y, *et al.* BMP7 Signaling in TGFBR2-Deficient Stromal Cells Provokes Epithelial Carcinogenesis. *Mol Cancer Res* 2018;16:1568–78.
- 46 Motoyama K, Tanaka F, Kosaka Y, *et al.* Clinical significance of BMP7 in human colorectal cancer. *Ann Surg Oncol* 2008;15:1530–7.
- 47 Bien SA, Su Y-R, Conti DV, *et al.* Genetic variant predictors of gene expression provide new insight into risk of colorectal cancer. *Hum Genet* 2019;138:307–26.
- 48 Favaro E, Bensaad K, Chong MG, *et al.* Glucose utilization via glycogen phosphorylase sustains proliferation and prevents premature senescence in cancer cells. *Cell Metab* 2012;16:751–64.
- 49 Jänne PA, Mayer RJ. Chemoprevention of colorectal cancer. *N Engl J Med* 2000;342:1960–8.
- 50 Goessling W, North TE, Loewer S, *et al.* Genetic interaction of PGE2 and Wnt signaling regulates developmental specification of stem cells and regeneration. *Cell* 2009;136:1136–47.
- 51 Castellone MD, Teramoto H, Williams BO, *et al.* Prostaglandin E2 promotes colon cancer cell growth through a Gs- α - β -catenin signaling axis. *Science* 2005;310:1504–10.
- 52 Shoji Y, Takahashi M, Kitamura T, *et al.* Downregulation of prostaglandin E receptor subtype EP3 during colon cancer development. *Gut* 2004;53:1151–8.
- 53 Tenesa A, Farrington SM, Prendergast JGD, *et al.* Genome-wide association scan identifies a colorectal cancer susceptibility locus on 11q23 and replicates risk loci at 8q24 and 18q21. *Nat Genet* 2008;40:631–7.
- 54 Tomlinson IPM, Webb E, Carvajal-Carmona L, *et al.* A genome-wide association study identifies colorectal cancer susceptibility loci on chromosomes 10p14 and 8q23.3. *Nat Genet* 2008;40:623–30.
- 55 Guinney J, Dienstmann R, Wang X, *et al.* The consensus molecular subtypes of colorectal cancer. *Nat Med* 2015;21:1350–6.
- 56 Tan J, McKenzie C, Potamitis M, *et al.* The role of short-chain fatty acids in health and disease. *Adv Immunol* 2014;121:91–119.
- 57 McNabney SM, Henagan 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.



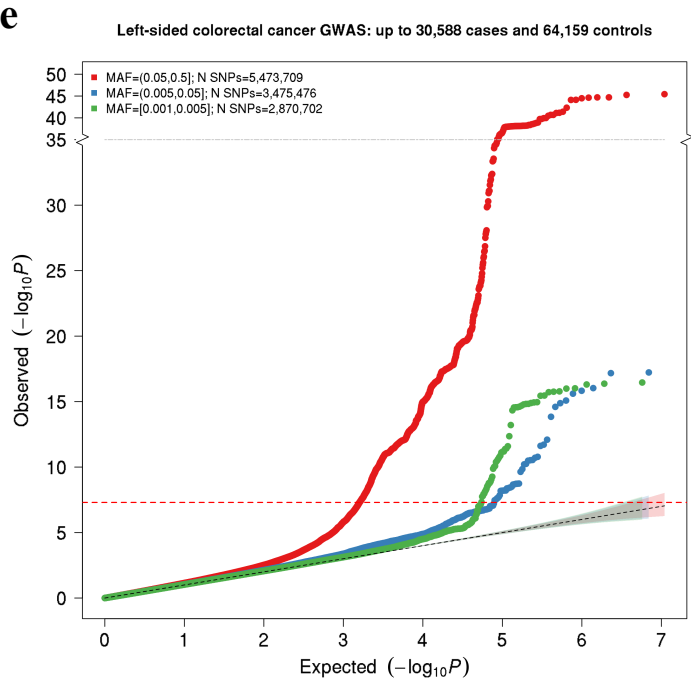
Supplementary figure 1. Distributions of age of diagnosis by primary tumor anatomic subsite and sex. Note that our data recapitulate the previously reported higher percentage of female proximal colon cancer cases, a male-to-female ratio that increases progressively from the proximal colon to the rectum, and differences in age of onset by primary tumor site, with an earlier age of onset for rectal cancer. The red dashed lines and blue dotted lines denote median and mean age of diagnosis within stratum, respectively.



Supplementary figure 2. Quantile-quantile (QQ) plots stratified by minor allele frequency (MAF) bins for the five GWAS meta-analyses of colorectal cancer case subgroups defined by primary tumor anatomic subsite. a, Proximal colon cancer GWAS. b, Distal colon cancer GWAS. c, Rectal cancer GWAS. d, Colon cancer GWAS. e, Left-sided colorectal cancer GWAS. GWAS studies were imputed to the Haplotype Reference Consortium (HRC) panel. The red dashed line indicates the genome-wide significance threshold ($P=5 \times 10^{-8}$). The transparent regions around the equality line represent the analytically estimated 95% confidence bands for each MAF bin.

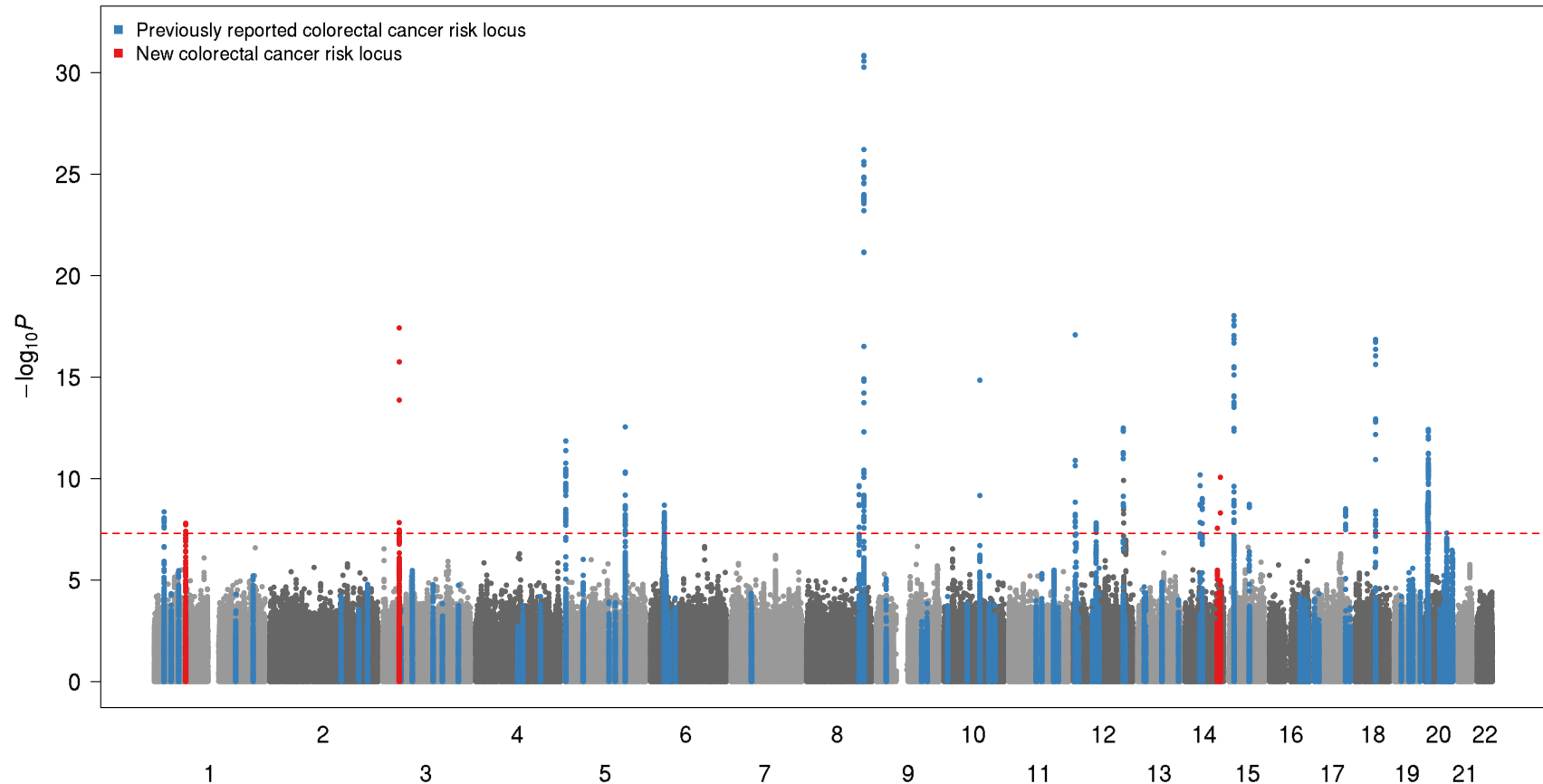


Supplementary figure 2. (continued)



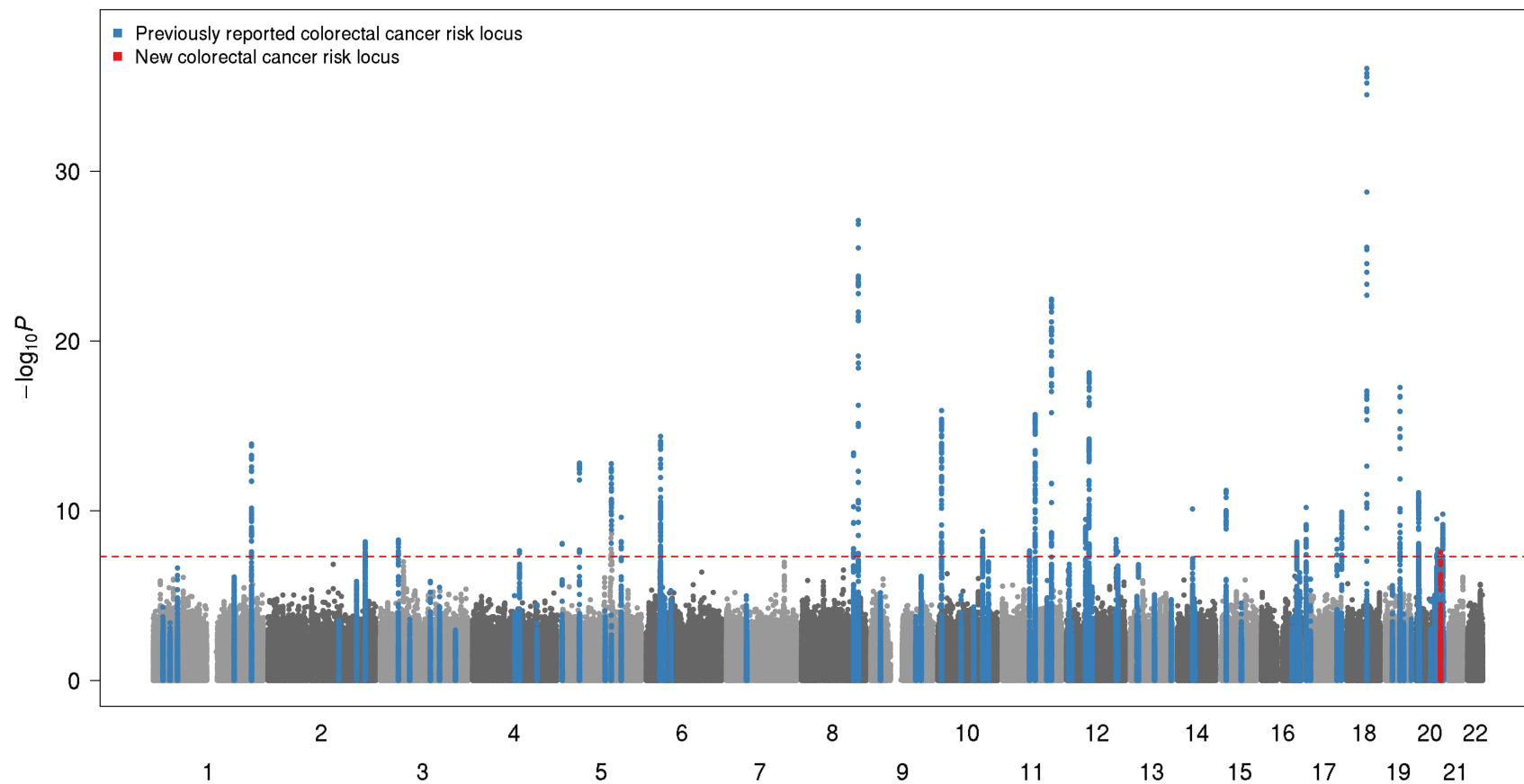
Supplementary figure 2. (continued)

Proximal colon cancer GWAS: up to 15,706 cases and 64,159 controls
11,437,803 variants



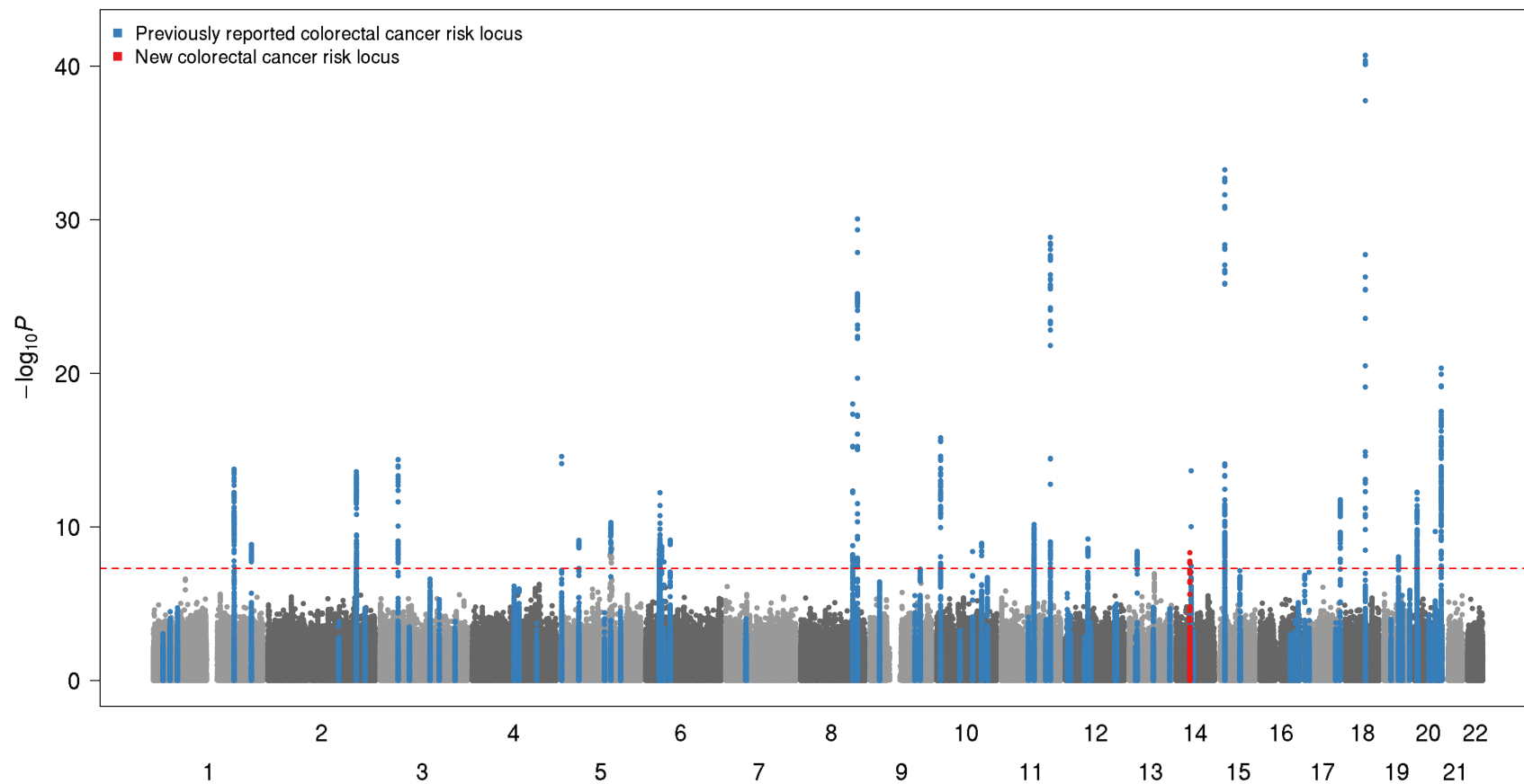
Supplementary figure 3. Manhattan plots showing results of the five GWAS meta-analyses of CRC case subgroups defined by primary tumor anatomic subsite. GWAS studies were imputed to the Haplotype Reference Consortium (HRC) panel. Association results for each variant ($-\log_{10} P$ values) are plotted against genomic position (NCBI Build 37). The red dashed line indicates the genome-wide significance threshold ($P=5 \times 10^{-8}$). New loci are shown in red in the Manhattan plot. Loci previously associated with overall colorectal cancer risk at genome-wide significance are denoted in blue. (figure continued on following pages)

Distal colon cancer GWAS: up to 14,376 cases and 64,159 controls
11,353,757 variants



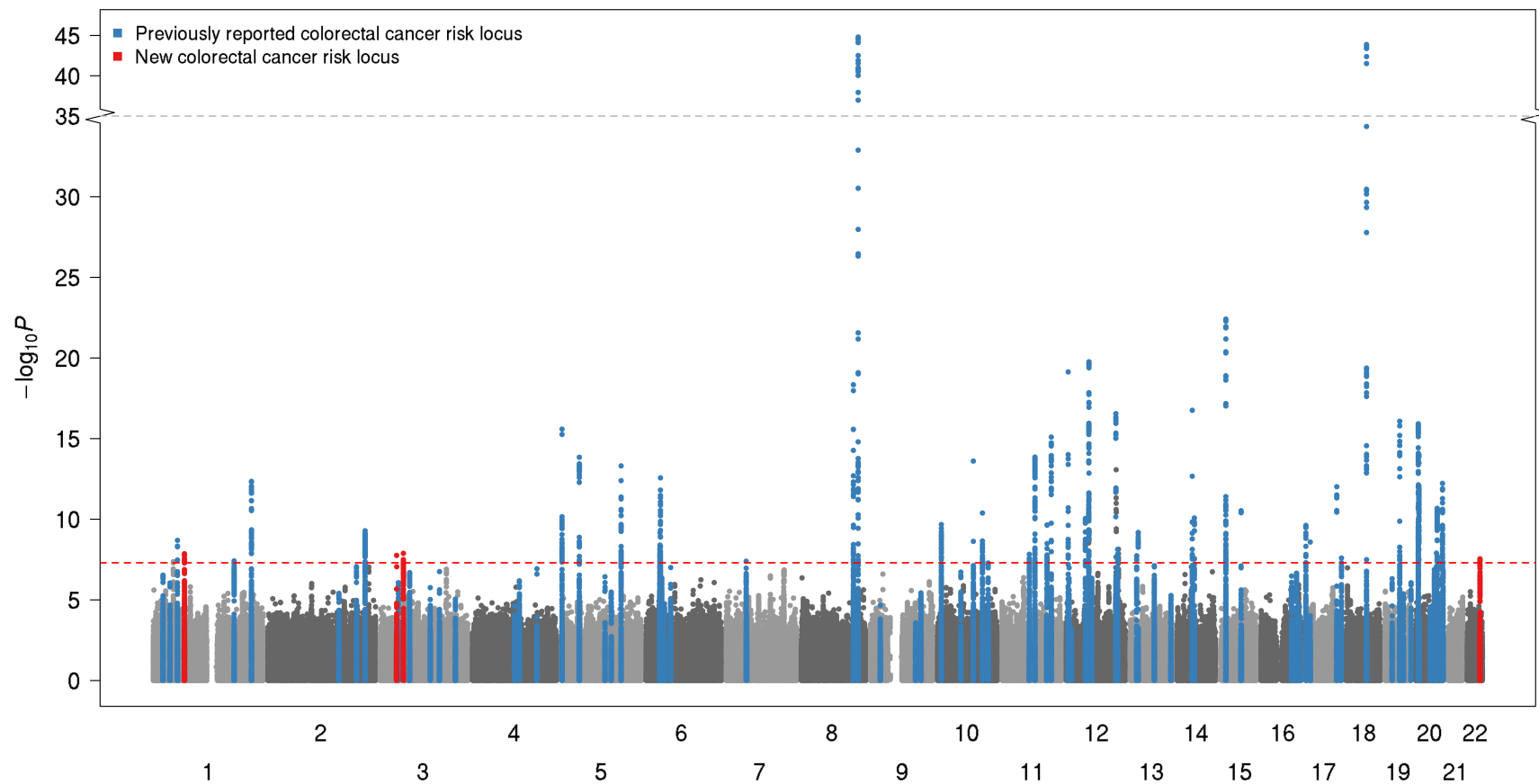
Supplementary figure 3. (continued)

Rectal cancer GWAS: up to 16,212 cases and 64,159 controls
11,358,342 variants



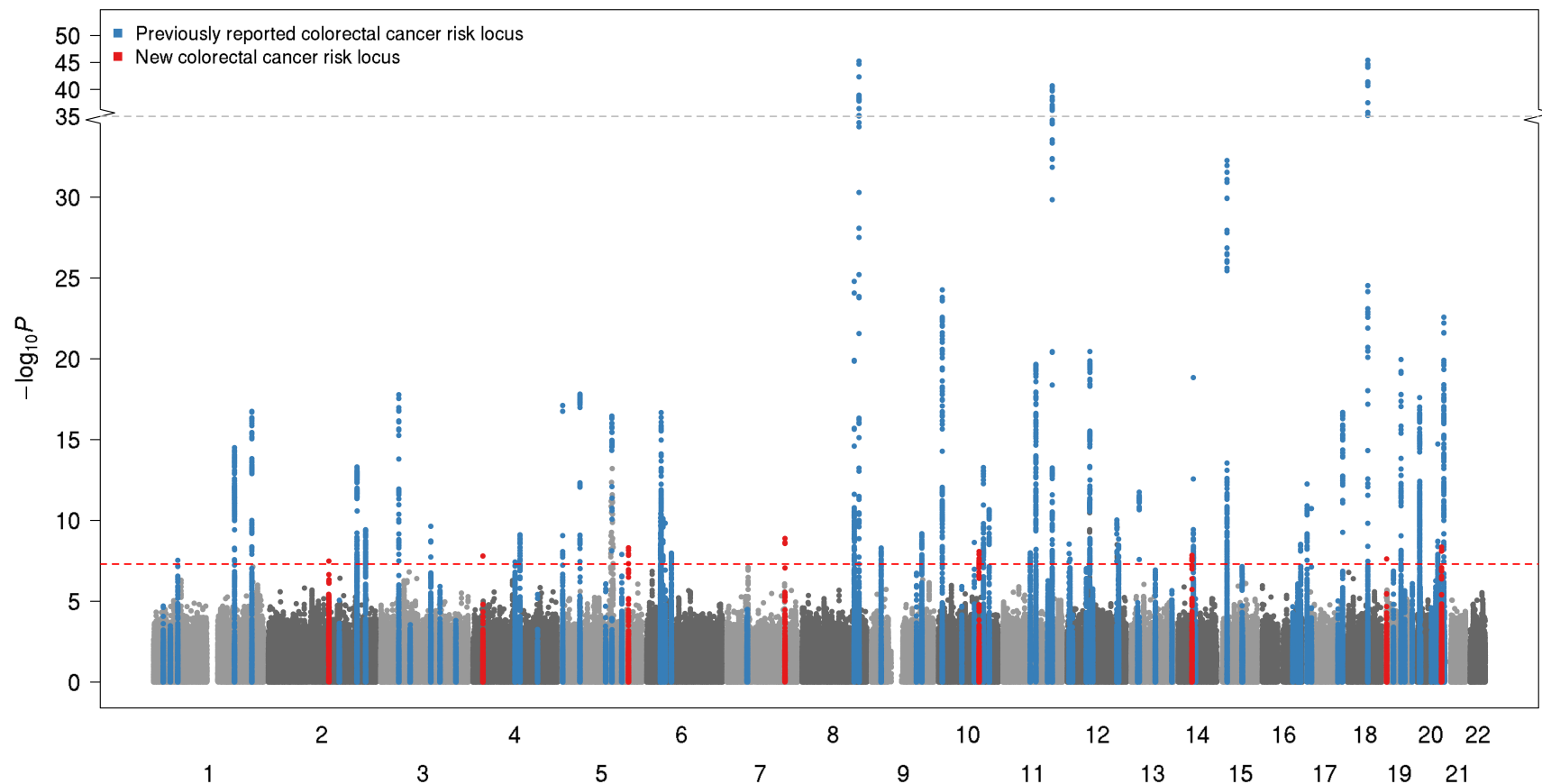
Supplementary figure 3. (continued)

Colon cancer GWAS: up to 32,002 cases and 64,159 controls
11,928,745 variants



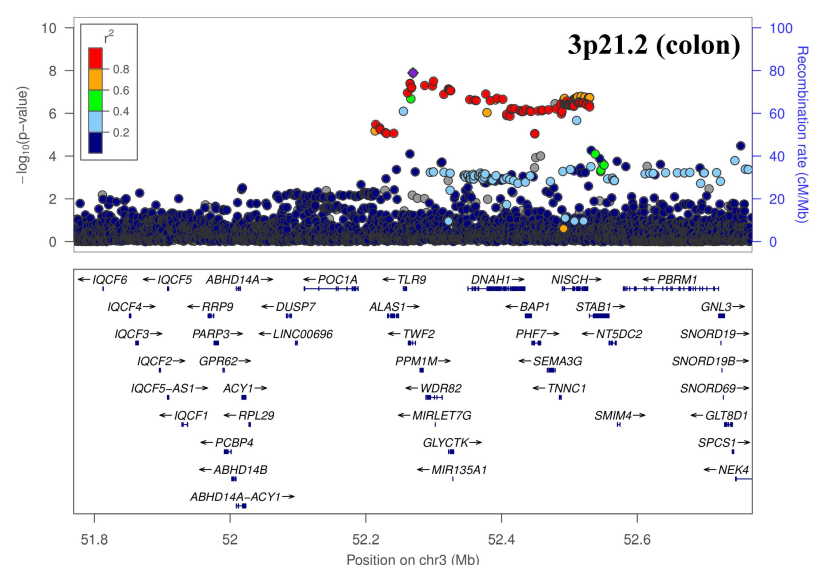
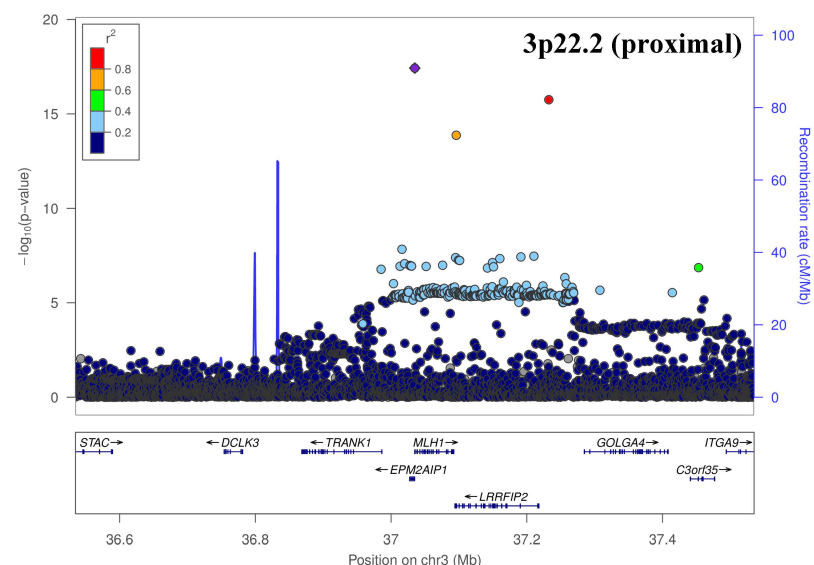
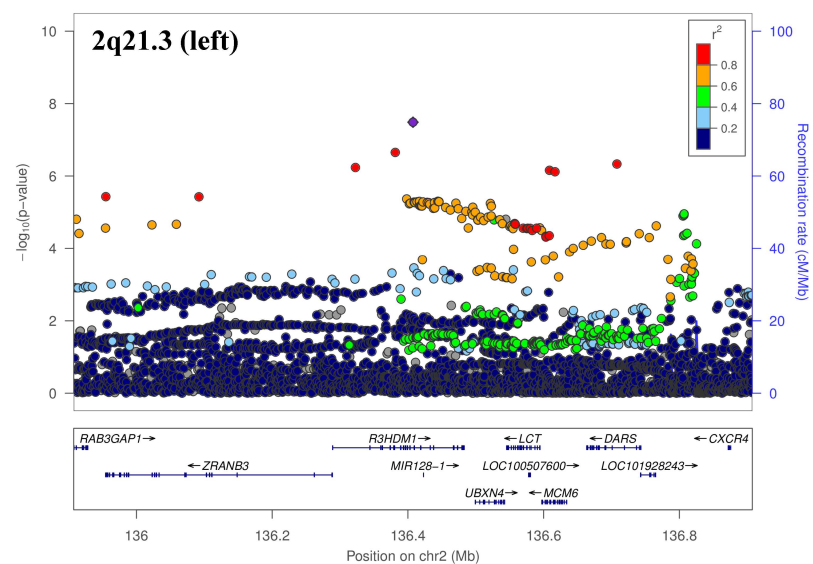
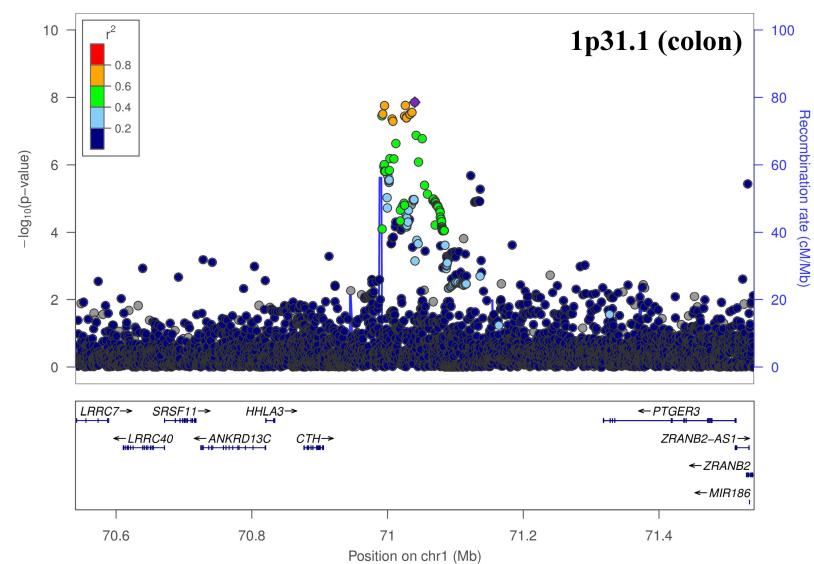
Supplementary figure 3. (continued)

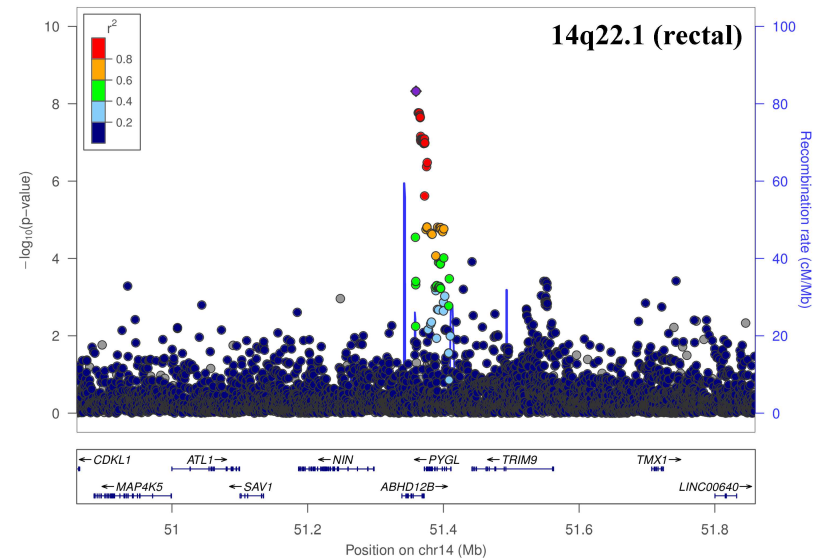
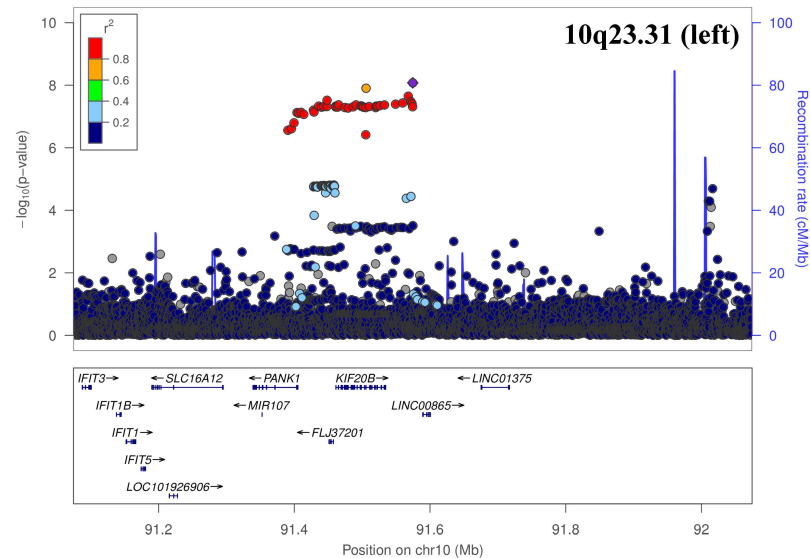
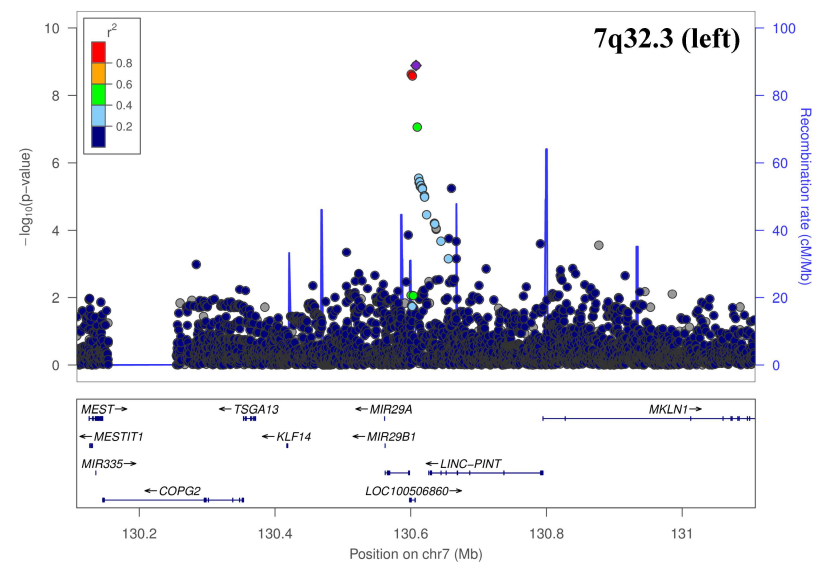
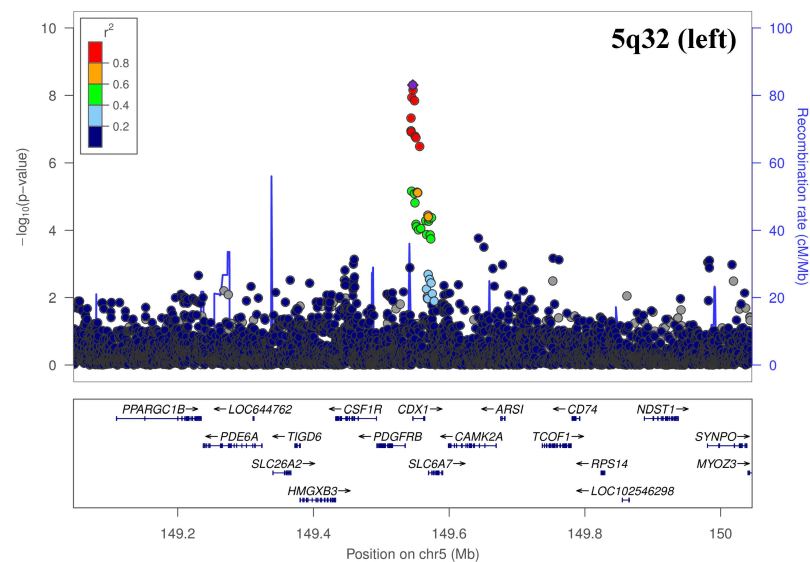
Left-sided colorectal cancer GWAS: up to 30,588 cases and 64,159 controls
11,819,887 variants

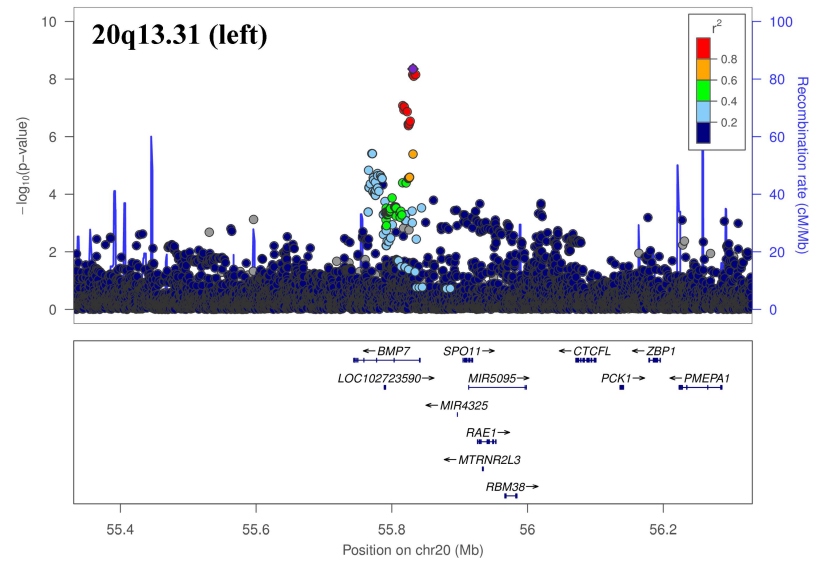
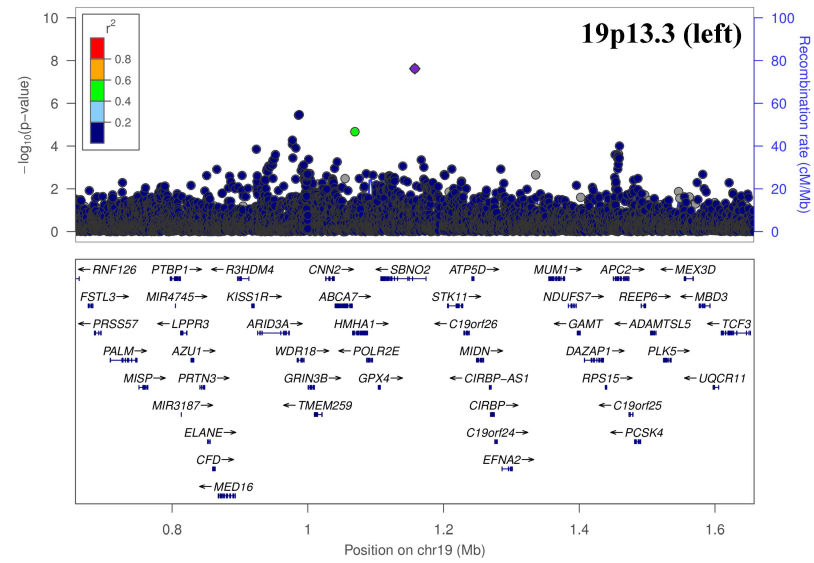
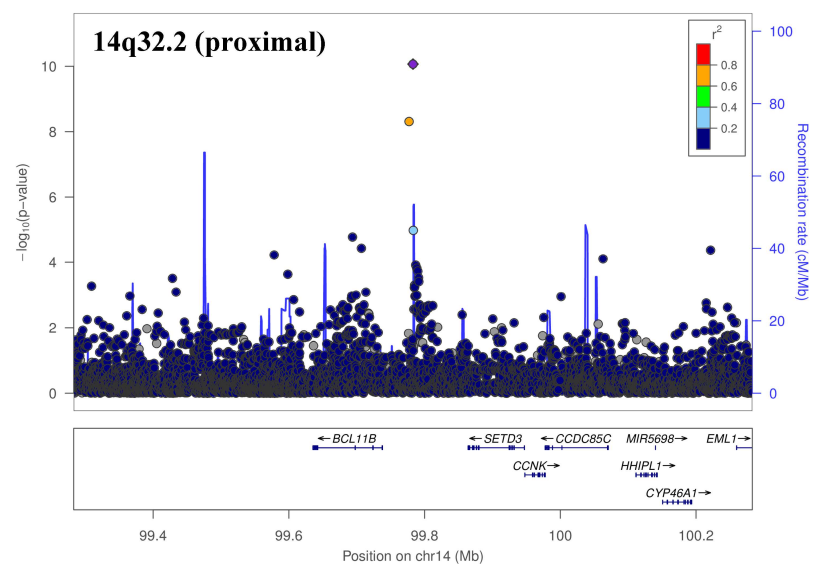
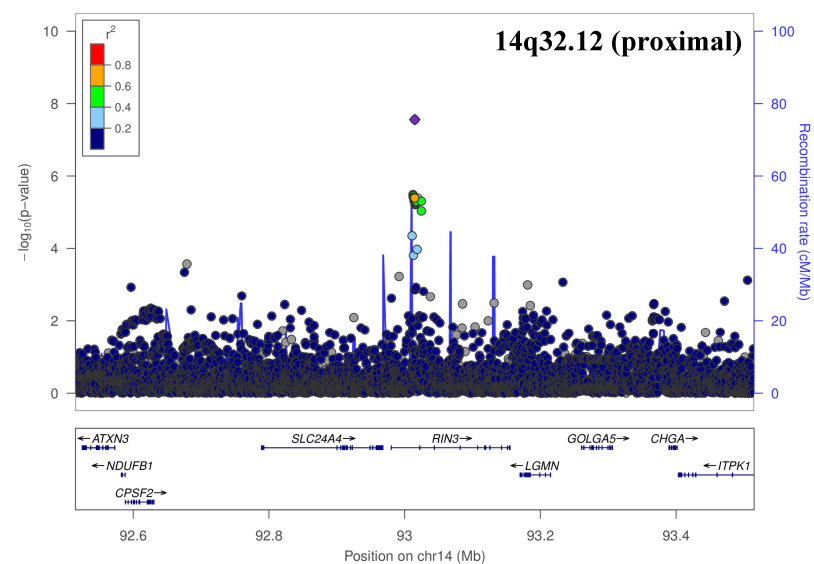


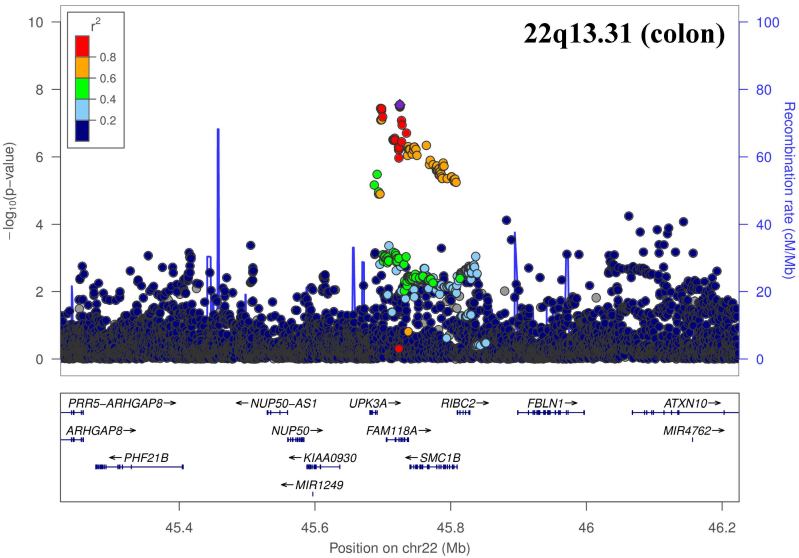
Supplementary figure 3. (continued)

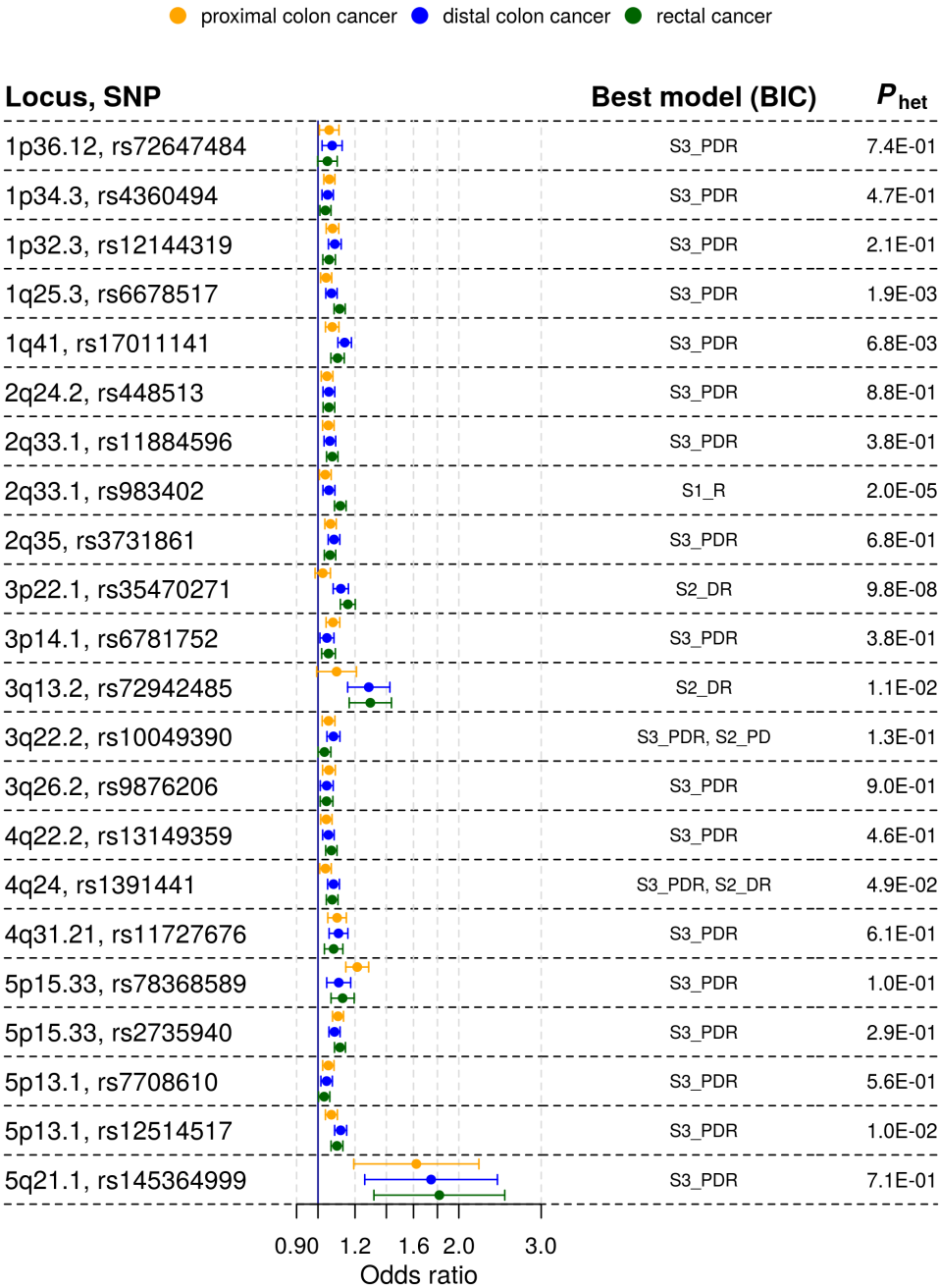
Supplementary figure 4. Regional association plots for the new CRC risk loci reaching genome-wide significance (P -value $< 5 \times 10^{-8}$) in the GWAS meta-analyses for CRC case subgroups defined by primary tumor anatomical subsites. Case subgroups were defined as follows: proximal colon cancer ($n=15,706$), distal colon cancer ($n=14,376$), rectal cancer ($n=16,212$), colon cancer ($n=32,002$), and distal/left-sided CRC ($n=30,588$). Analyses were based on 64,159 shared controls. LocusZoom plots show the $-\log_{10}(P\text{-value})$ for the association with risk for the CRC case subgroup as a function of genomic position (NCBI Build 37) for each variant within a 1-Mb window centered at the lead variant of the locus. Lead variants are indicated by the purple diamond symbol. The color labeling of other variants indicates LD with the lead variant estimated from our previously published whole-genome sequence (WGS) data on 2,159 European ancestry study participants (Huyghe *et al.*). Gray dots indicate that the variant was not found in our WGS panel and that LD could not be calculated. Recombination rates are based on Phase 2 HapMap and gene models are RefSeq genes taken from the UCSC Genome Browser.



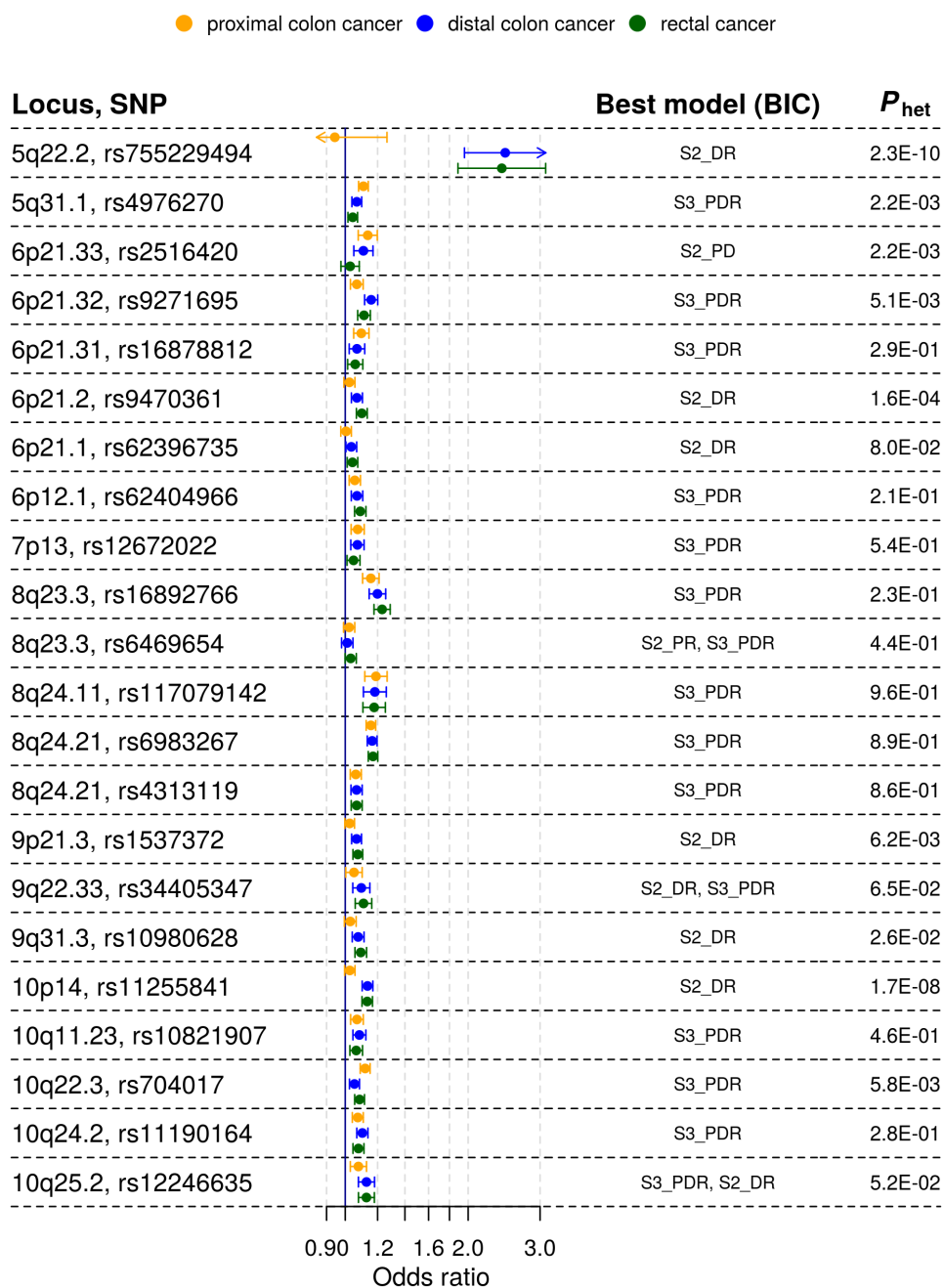




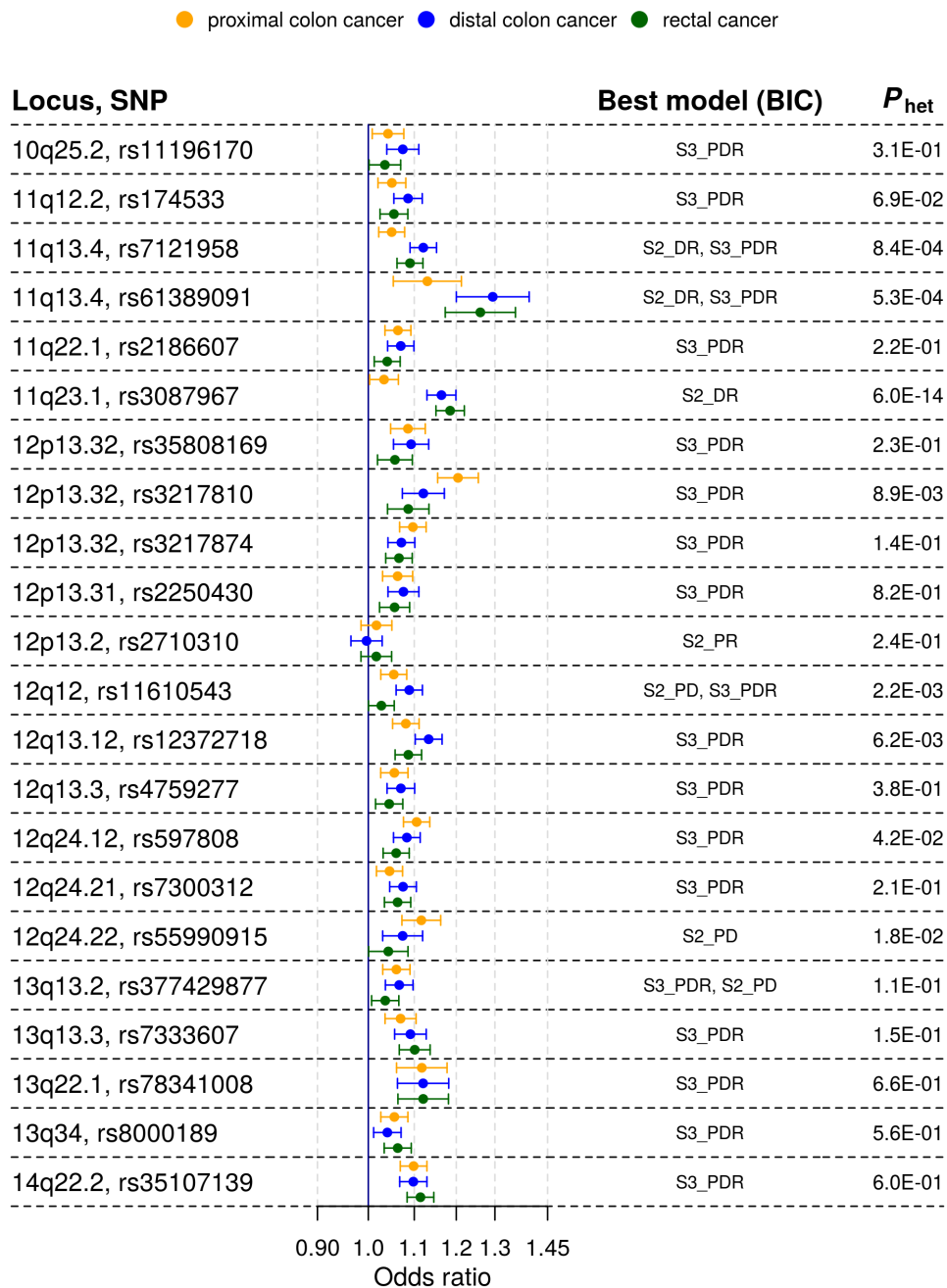




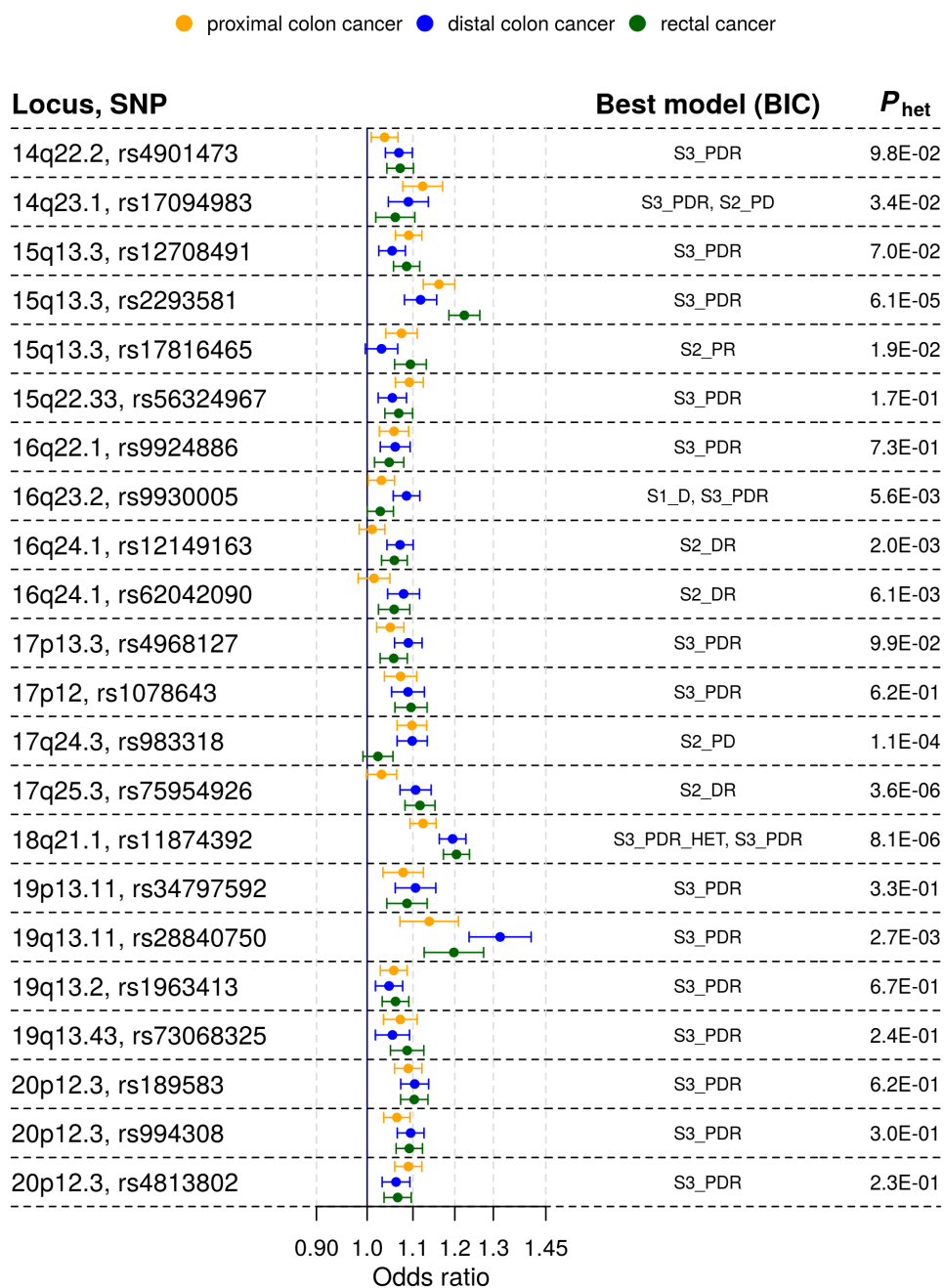
Supplementary figure 5. Forest plots and multinomial modeling results for previously reported CRC risk variants. Best model is the best-fitting multinomial logistic regression model according to the Bayesian Information Criterion (BIC). Please refer to supplementary table 2 for model definitions. P_{het} is the P -value from a heterogeneity test, testing the null hypothesis that odds ratios are fixed across CRC subtypes defined by primary tumor site.



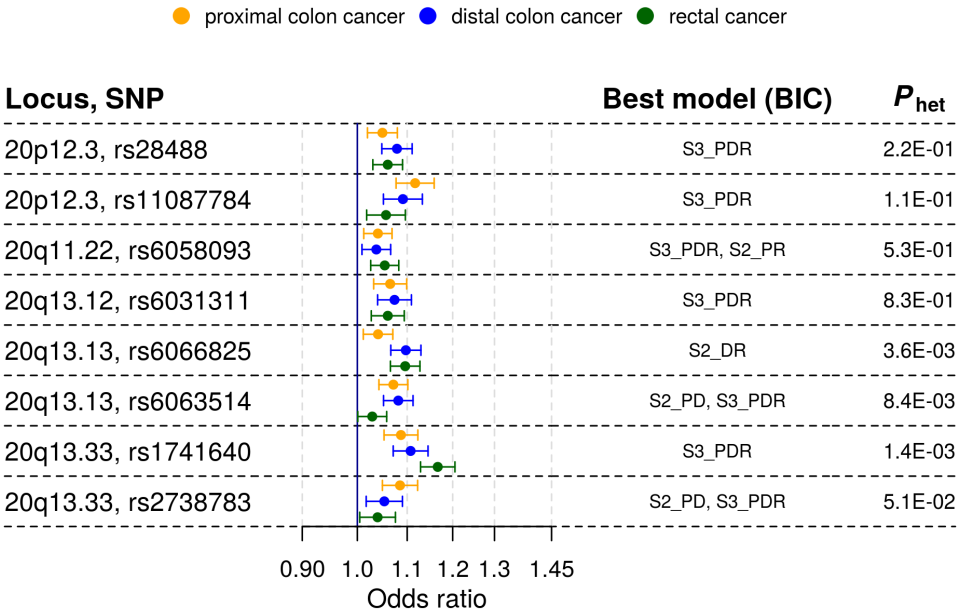
Supplementary figure 5 (continued).



Supplementary figure 5 (continued).



Supplementary figure 5 (continued).



Supplementary figure 5 (continued).

Supplementary Table 4. Most likely target gene(s) at the 13 new loci identified across the CRC case subgroup analyses.

Locus	Putative target gene(s)	Expression-based linking (GTEx V8)	Biological relevance, experimental functional evidence, somatic alterations, familial syndromes
1p31.1	<i>PTGER3</i>	<i>PTGER3</i> (7 tissues)	<i>PTGER3</i> encodes Prostaglandin E Receptor 3, a receptor for prostaglandin E2 (PGE2), a potent pro-inflammatory metabolite that is biosynthesized by Cyclooxygenase-2 (COX-2). COX-2 plays a critical role in mediating inflammatory responses that lead to epithelial malignancies and its expression is induced by NF- κ B and TNF- α . The anti-inflammatory activity of nonsteroidal anti-inflammatory drugs (NSAIDs) such as aspirin and ibuprofen operates mainly through COX-2 inhibition, and long-term NSAID use decreases incidence and mortality from CRC.[1] Prostaglandin E2 (PGE2) is required for the activation of β -catenin by Wnt in stem cells,[2] and promotes colon cancer cell growth.[3] Prostaglandin E Receptor 3 plays an important role in suppression of cell growth and its downregulation was shown to enhance colon carcinogenesis.[4] Hypermethylation may contribute to its downregulation in colon cancer.[4]
2q21.3	<i>LCT</i>		Lead SNP rs1446585 is in strong LD with the functional SNP rs4988235 (LD $r^2 = 0.854$) in the <i>cis</i> -regulatory element of the lactase gene. In Europeans, the rs4988235 genotype determines the autosomal dominant lactase persistence phenotype, or the ability to digest the milk sugar lactose in adulthood. The allele determining lactase persistence (T) is associated with a decreased risk of CRC. This is consistent with a previous candidate study that reported a significant association between low lactase activity defined by the CC genotype and CRC risk in the Finnish population.[5] The protective effect conferred

			by the lactase persistence genotype is likely mediated by dairy products and calcium which are known protective factors for CRC.[6] Consistent with a dominant model, associations for rs1446585 and rs4988235 became more significant when tested assuming a dominant model with P -values of 4.4×10^{-11} and 1.4×10^{-9} , respectively (see main text).
3p22.2	<i>MLH1</i>	<i>MLH1</i> (14 tissues)	Previous candidate gene studies have reported strong and robust associations between the common, <i>MLH1</i> gene promoter region and lead SNP rs1800734, and sporadic CRC cases with high microsatellite instability (MSI-H) status with consistent direction of effects.[7,8] Rare deleterious nonsynonymous mutations in the DNA mismatch repair (MMR) gene <i>MLH1</i> are a cause of Lynch syndrome (OMIM #609310). The risk allele of the likely causal SNP rs1800734 showed a strong association with <i>MLH1</i> promoter hypermethylation and loss of MLH1 protein in CRC tumors.[8] The mechanisms of <i>MLH1</i> promoter hypermethylation and subsequent gene silencing may account for most sporadic CRC tumors with defective DNA MMR and MSI-H.[9]
3p21.2	<i>STAB1</i> ; <i>TLR9</i> ; <i>NISCH</i>	<i>STAB1</i> (10 tissues); <i>TLR9</i> (3 tissues); <i>NISCH</i> (4 tissues)	This signal is located in a gene dense region. The Stabilin 1 (<i>STAB1</i>) gene encodes an endocytotic scavenger receptor expressed in a number of cell types, including activated macrophages in human malignancies.[10] A rare missense variant in <i>STAB1</i> has previously shown to be strongly associated with serum lactate dehydrogenase (LDH) levels,[11] a widely used marker of tissue damage, affirming a link between <i>STAB1</i> and the clearance of products of cell lysis through the mononuclear phagocytic system. Human Protein Atlas data based on The Cancer Genome Atlas (TCGA) show that <i>STAB1</i> expression is an unfavorable prognostic marker for CRC (logrank test $P=0.0008$, based on maximally separated Kaplan-Meier curves; n samples=597). Lead SNP rs353548 is

			located in an intron of the toll like receptor 9 (<i>TLR9</i>) gene which could also be involved. This key component of innate and adaptive immunity is a drug target for many immune-mediated diseases, and the antagonist drug hydroxychloroquine is included in chemotherapy combination clinical trials for colorectal carcinoma (ClinicalTrials.gov Identifier: NCT01006369). The Nischarin (<i>NISCH</i>) gene encodes an $\alpha 5$ integrin-binding protein and may be a tumor suppressor gene that limits breast cancer progression.[12] Nischarin inhibits Rac-induced cell migration and invasion in breast and colon epithelial cells.[13]
5q32	<i>CDX1</i>		The intestine-specific transcription factor caudal-type homeobox 1(<i>CDX1</i>) 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. Colonic crypt cells express BMI1 but not CDX1. The repression of BMI1 is mediated by microRNA-215 which acts as a target of CDX1 to promote differentiation and inhibit stemness.[14] Consistent with this view, CDX1 has been shown to inhibit human colon cancer cell proliferation by blocking β -catenin/T-cell factor transcriptional activity.[15]
7q32.3	<i>KLF14</i> ; <i>LINC00513</i>	<i>LINC00513</i> (transverse colon + 2 tissues)	The Krüppel-like factor 14 (<i>KLF14</i>) gene is a strong candidate involved in TGF- β signaling. We previously reported loci at known CRC oncogene <i>KLF5</i> and at <i>KLF2</i> . [16] The imprinted gene <i>KLF14</i> shows monoallelic maternal expression, and is induced by TGF- β to transcriptionally corepress the TGF-beta receptor II (<i>TGFBR2</i>) gene.[17] A cis-eQTL for <i>KLF14</i> , that is uncorrelated with our lead SNP rs73161913, acts as a master

			regulator related to multiple metabolic phenotypes,[18,19] and an independent variant in this region has been associated to basal cell carcinoma.[20] The signal overlaps with an eQTL for the lncRNA gene <i>LINC00513</i> which may be involved in the regulation of <i>KLF14</i> expression.
10q23.31	<i>PANK1</i> ; <i>KIF20B</i>	<i>PANK1</i> (transverse colon + 3 tissues); <i>KIF20B</i> (transverse colon + 7 tissues)	At 10q23.31, GTEx data show that the lead SNP rs7071258 is an eQTL in transverse colon tissue for genes Pantothenate Kinase 1 (<i>PANK1</i>) and Kinesin Family Member 20B (<i>KIF20B</i>). The enzyme encoded by <i>PANK1</i> catalyzes the rate-limiting reaction in the biosynthesis of coenzyme A and may play a role in tumor metabolism.[21] <i>KIF20B</i> has been suggested to play an oncogenic role in bladder carcinogenesis.[22] <i>KIF20B</i> missense variant rs34354493 (canonical transcript, p.Lys1609Glu) is in high LD with the lead variant ($r^2=0.90$) and is predicted to be deleterious by multiple algorithms (CADD, DANN, Polyphen, SIFT).
14q22.1	<i>PYGL</i> ; <i>NIN</i> ; <i>ABHD12B</i>	<i>PYGL</i> (transverse colon + 12 tissues); <i>ABHD12B</i> (transverse colon + 8 tissues); <i>NIN</i> (transverse	GTEx data show that, in gastrointestinal tissues, the lead SNP is a <i>cis</i> -eQTL co-regulating expression of genes <i>PYGL</i> , <i>ABHD12B</i> , and <i>NIN</i> . Glycogen Phosphorylase L (<i>PYGL</i>) is the strongest candidate. We recently identified and replicated an association between genetically predicted <i>PYGL</i> expression and CRC risk in a transcriptome-wide association study that used transverse colon tissue transcriptomes and genotypes from GTEx to construct prediction models.[23] Favaro <i>et al.</i> showed that this glycogen metabolism gene plays an important role in sustaining proliferation and preventing premature senescence in hypoxic cancer cells.[24] In different cancer cells lines, silencing of <i>PYGL</i> , expression of which is induced by exposure to hypoxia, led to increased glycogen accumulation and

		colon + 2 tissues)	increased reactive oxygen species levels that contributed to p53-dependent induction of senescence and impaired tumorigenesis.[24]
14q32.12	<i>RIN3</i>	<i>RIN3</i> (transverse colon + 11 tissues)	Lead SNP rs61975764 is an eQTL for gene Ras And Rab Interactor 3 (<i>RIN3</i>) in colon tissue, the risk allele G being associated with decreased expression. <i>RIN3</i> functions as a RAB5 and RAB31 guanine nucleotide exchange factor involved in endocytosis.[25,26]
14q32.2	<i>BCL11B</i>		The lead SNP rs80158569 of this highly localized proximal colon-specific association signal is located in a normal colonic crypt enhancer region and overlaps with multiple transcription factor binding sites, making it a strong functional candidate. The nearby gene <i>BCL11B</i> encodes a transcription factor that is required for normal T cell development,[27,28] and that has been identified as a SWI/SNF complex subunit.[29] <i>BCL11B</i> acts as a haploinsufficient tumor suppressor in T-cell acute lymphoblastic leukemia (T-ALL).[30,31] Experimental work reported by Sakamaki <i>et al.</i> suggests that impairment of Bcl11b promotes intestinal tumorigenesis in mice and humans through deregulation of the β -catenin pathway.[32]
19p13.3	<i>STK11</i> ; <i>SBNO2</i>		This signal is located in a gene-dense region. Lead SNP rs62131228 is intronic to gene Strawberry notch homologue 2 (<i>SBNO2</i>), a transcriptional corepressor of NF- κ B in macrophages that plays a role in the STAT3-regulated anti-inflammatory signaling pathway.[33] The nearby tumor suppressor gene Serine/Threonine Kinase 11 (<i>STK11</i>) is an especially plausible candidate effector gene. Mutations in this gene cause Peutz-Jeghers syndrome (OMIM #175200), an autosomal dominant disorder characterized by the growth

			of hamartomatous gastrointestinal polyps and an increased risk of various neoplasms.[34,35]
20q13.31	<i>BMP7</i>	<i>BMP7</i> (3 tissues)	The Bone Morphogenetic Protein 7 (BMP7) gene is a strong candidate. In normal intestinal cell crypts, various gradients of TGF- β family members interact with the antagonistic Wnt signaling pathway to maintain homeostasis. Members of the TGF- β family, including several bone morphogenetic proteins (BMPs), frequently have somatic mutations in sporadic CRC tumors, have been implicated by GWASs, and germline mutations are causative for familial CRC syndromes.[36] BMP7 signaling in <i>TGFBR2</i> -deficient stromal cells promotes epithelial carcinogenesis through SMAD4-mediated signaling.[37] In CRC tumors, <i>BMP7</i> expression correlates with parameters of pathological aggressiveness such as liver metastasis and poor prognosis.[38]
22q13.31	<i>FAM118A</i> ; <i>FBLN1</i>	<i>FAM118A</i> (transverse colon + 40 tissues)	GTEx data show that the lead SNP rs736037 is an eQTL for gene <i>FAM118A</i> in many tissues, including transverse colon. The function of <i>FAM118A</i> is poorly understood. <i>FAM118A</i> missense variant rs6007594 (canonical transcript, p.Arg239His) is in high LD with the lead variant rs736037 ($r^2=0.96$) and is predicted to be deleterious by multiple algorithms (CADD, DANN, Polyphen). The protein encoded by the nearby Fibulin 1 (<i>FBLN1</i>) gene plays a role in the organization and function of the extracellular matrix and basement membranes. <i>FBLN1</i> has been implicated in tumor-related processes and both oncogenic and tumor-suppressive properties have been described for this protein.[39] Other genes in the region are no obvious candidates.

References

- 1 Jänne PA, Mayer RJ. Chemoprevention of colorectal cancer. *N Engl J Med* 2000;**342**:1960–8. doi:10.1056/NEJM200006293422606
- 2 Goessling W, North TE, Loewer S, *et al.* Genetic interaction of PGE2 and Wnt signaling regulates developmental specification of stem cells and regeneration. *Cell* 2009;**136**:1136–47. doi:10.1016/j.cell.2009.01.015
- 3 Castellone MD, Teramoto H, Williams BO, *et al.* Prostaglandin E2 promotes colon cancer cell growth through a Gs-axin-beta-catenin signaling axis. *Science* 2005;**310**:1504–10. doi:10.1126/science.1116221
- 4 Shoji Y, Takahashi M, Kitamura T, *et al.* Downregulation of prostaglandin E receptor subtype EP3 during colon cancer development. *Gut* 2004;**53**:1151–8. doi:10.1136/gut.2003.028787
- 5 Rasinperä H, Forsblom C, Enattah NS, *et al.* The C/C-13910 genotype of adult-type hypolactasia is associated with an increased risk of colorectal cancer in the Finnish population. *Gut* 2005;**54**:643–7. doi:10.1136/gut.2004.055939
- 6 World Cancer Research Fund/American Institute for Cancer Research. Continuous Update Project Expert Report 2018. Diet, nutrition, physical activity and colorectal cancer. Available at dietandcancerreport.org
- 7 Raptis S, Mrkonjic M, Green RC, *et al.* MLH1 -93G>A promoter polymorphism and the risk of microsatellite-unstable colorectal cancer. *J Natl Cancer Inst* 2007;**99**:463–74. doi:10.1093/jnci/djk095
- 8 Mrkonjic M, Roslin NM, Greenwood CM, *et al.* Specific variants in the MLH1 gene region may drive DNA methylation, loss of protein expression, and MSI-H colorectal cancer. *PLoS One* 2010;**5**:e13314. doi:10.1371/journal.pone.0013314
- 9 Cunningham JM, Christensen ER, Tester DJ, *et al.* Hypermethylation of the hMLH1 promoter in colon cancer with microsatellite instability. *Cancer Res* 1998;**58**:3455–60.
- 10 Schönhaar K, Schledzewski K, Michel J, *et al.* Expression of stabilin-1 in M2 macrophages in human granulomatous disease and melanocytic lesions. *Int J Clin Exp Pathol* 2014;**7**:1625–34.
- 11 Kristjansson RP, Oddsson A, Helgason H, *et al.* Common and rare variants associating with serum levels of creatine kinase and lactate dehydrogenase. *Nat Commun* 2016;**7**:10572. doi:10.1038/ncomms10572
- 12 Baranwal S, Wang Y, Rathinam R, *et al.* Molecular characterization of the tumor-suppressive function of nischarin in breast cancer. *J Natl Cancer Inst* 2011;**103**:1513–28. doi:10.1093/jnci/djr350
- 13 Alahari SK. Nischarin inhibits Rac induced migration and invasion of epithelial cells by affecting signaling cascades involving PAK. *Exp Cell Res* 2003;**288**:415–24.
- 14 Jones MF, Hara T, Francis P, *et al.* The CDX1-microRNA-215 axis regulates colorectal cancer stem cell differentiation. *Proc Natl Acad Sci USA* 2015;**112**:E1550–8. doi:10.1073/pnas.1503370112
- 15 Guo R-J, Huang E, Ezaki T, *et al.* Cdx1 inhibits human colon cancer cell proliferation by reducing beta-catenin/T-cell factor transcriptional activity. *J Biol Chem* 2004;**279**:36865–75. doi:10.1074/jbc.M405213200
- 16 Huyghe JR, Bien SA, Harrison TA, *et al.* Discovery of common and rare genetic risk variants for colorectal cancer. *Nat Genet*

- 2019;**51**:76–87. doi:10.1038/s41588-018-0286-6
- 17 Truty MJ, Lomberg G, Fernandez-Zapico ME, *et al.* Silencing of the transforming growth factor-beta (TGFbeta) receptor II by Kruppel-like factor 14 underscores the importance of a negative feedback mechanism in TGFbeta signaling. *J Biol Chem* 2009;**284**:6291–300. doi:10.1074/jbc.M807791200
 - 18 Small KS, Hedman AK, Grundberg E, *et al.* Identification of an imprinted master trans regulator at the KLF14 locus related to multiple metabolic phenotypes. *Nat Genet* 2011;**43**:561–4. doi:10.1038/ng.833
 - 19 Small KS, Todorčević M, Civelek M, *et al.* Regulatory variants at KLF14 influence type 2 diabetes risk via a female-specific effect on adipocyte size and body composition. *Nat Genet* 2018;**50**:572–80. doi:10.1038/s41588-018-0088-x
 - 20 Stacey SN, Sulem P, Masson G, *et al.* New common variants affecting susceptibility to basal cell carcinoma. *Nat Genet* 2009;**41**:909–14. doi:10.1038/ng.412
 - 21 Wang S-J, Yu G, Jiang L, *et al.* p53-Dependent regulation of metabolic function through transcriptional activation of pantothenate kinase-1 gene. *Cell Cycle* 2013;**12**:753–61. doi:10.4161/cc.23597
 - 22 Kanehira M, Katagiri T, Shimo A, *et al.* Oncogenic role of MPHOSPH1, a cancer-testis antigen specific to human bladder cancer. *Cancer Res* 2007;**67**:3276–85. doi:10.1158/0008-5472.CAN-06-3748
 - 23 Bien SA, Su Y-R, Conti DV, *et al.* Genetic variant predictors of gene expression provide new insight into risk of colorectal cancer. *Hum Genet* 2019;**138**:307–26. doi:10.1007/s00439-019-01989-8
 - 24 Favaro E, Bensaad K, Chong MG, *et al.* Glucose utilization via glycogen phosphorylase sustains proliferation and prevents premature senescence in cancer cells. *Cell Metab* 2012;**16**:751–64. doi:10.1016/j.cmet.2012.10.017
 - 25 Kajiho H, Saito K, Tsujita K, *et al.* RIN3: a novel Rab5 GEF interacting with amphiphysin II involved in the early endocytic pathway. *J Cell Sci* 2003;**116**:4159–68. doi:10.1242/jcs.00718
 - 26 Kajiho H, Sakurai K, Minoda T, *et al.* Characterization of RIN3 as a guanine nucleotide exchange factor for the Rab5 subfamily GTPase Rab31. *J Biol Chem* 2011;**286**:24364–73. doi:10.1074/jbc.M110.172445
 - 27 Avram D, Califano D. The multifaceted roles of Bcl11b in thymic and peripheral T cells: impact on immune diseases. *J Immunol* 2014;**193**:2059–65. doi:10.4049/jimmunol.1400930
 - 28 Punwani D, Zhang Y, Yu J, *et al.* Multisystem Anomalies in Severe Combined Immunodeficiency with Mutant BCL11B. *N Engl J Med* 2016;**375**:2165–76. doi:10.1056/NEJMoa1509164
 - 29 Kadoch C, Hargreaves DC, Hodges C, *et al.* Proteomic and bioinformatic analysis of mammalian SWI/SNF complexes identifies extensive roles in human malignancy. *Nat Genet* 2013;**45**:592–601. doi:10.1038/ng.2628
 - 30 Gutierrez A, Kentsis A, Sanda T, *et al.* The BCL11B tumor suppressor is mutated across the major molecular subtypes of T-cell acute lymphoblastic leukemia. *Blood* 2011;**118**:4169–73. doi:10.1182/blood-2010-11-318873
 - 31 Neumann M, Vosberg S, Schlee C, *et al.* Mutational spectrum of adult T-ALL. *Oncotarget* 2015;**6**:2754–66. doi:10.18632/oncotarget.2218
 - 32 Sakamaki A, Katsuragi Y, Otsuka K, *et al.* Bcl11b SWI/SNF-complex subunit modulates intestinal adenoma and regeneration

- after γ -irradiation through Wnt/ β -catenin pathway. *Carcinogenesis* 2015;**36**:622–31. doi:10.1093/carcin/bgv044
- 33 El Kasmi KC, Smith AM, Williams L, *et al.* Cutting edge: A transcriptional repressor and corepressor induced by the STAT3-regulated anti-inflammatory signaling pathway. *J Immunol* 2007;**179**:7215–9. doi:10.4049/jimmunol.179.11.7215
- 34 Hemminki A, Tomlinson I, Markie D, *et al.* Localization of a susceptibility locus for Peutz-Jeghers syndrome to 19p using comparative genomic hybridization and targeted linkage analysis. *Nat Genet* 1997;**15**:87–90. doi:10.1038/ng0197-87
- 35 Hemminki A, Markie D, Tomlinson I, *et al.* A serine/threonine kinase gene defective in Peutz-Jeghers syndrome. *Nature* 1998;**391**:184–7. doi:10.1038/34432
- 36 Jung B, Staudacher JJ, Beauchamp D. Transforming Growth Factor β Superfamily Signaling in Development of Colorectal Cancer. *Gastroenterology* 2017;**152**:36–52. doi:10.1053/j.gastro.2016.10.015
- 37 Eikesdal HP, Becker LM, Teng Y, *et al.* BMP7 Signaling in TGFB2-Deficient Stromal Cells Provokes Epithelial Carcinogenesis. *Mol Cancer Res* 2018;**16**:1568–78. doi:10.1158/1541-7786.MCR-18-0120
- 38 Motoyama K, Tanaka F, Kosaka Y, *et al.* Clinical significance of BMP7 in human colorectal cancer. *Ann Surg Oncol* 2008;**15**:1530–7. doi:10.1245/s10434-007-9746-4
- 39 Fontanil T, Mohamedi Y, Cobo T, *et al.* Novel associations within the tumor microenvironment: fibulins meet adamts. *Front Oncol* 2019;**9**:796. doi:10.3389/fonc.2019.00796

SUPPLEMENTARY TEXT

Colorectal tumor anatomic sublocation definitions

We defined proximal colon cancer as any primary tumor arising in the cecum, ascending colon, hepatic flexure, or transverse colon (ICD-9 codes: 153.4, 153.6, 153.0, or 153.1, respectively); distal colon cancer as any primary tumor arising in the splenic flexure, descending colon, or sigmoid colon (ICD-9 codes: 153.7, 153.2, or 153.3, respectively); and rectal cancer as any primary tumor arising in the rectum or rectosigmoid junction (ICD-9 codes: 154.1, or 154.0, respectively). To examine the effect of including rectosigmoid junction in the rectal cancer analyses, we performed a sensitivity analysis in which we excluded rectosigmoid junction cases. Since we found that association signals were attenuated in the analyses without the rectosigmoid junction, we included it in the definition of rectal cancer. For the GWAS discovery analyses, we analyzed five case subgroups based on primary tumor 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 ICD-9 code 153.9. 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. Supplementary figure 1 and supplementary table 1 summarize distributions of age of diagnosis by sex and primary tumor site. Our data recapitulate the reported higher percentage of female proximal colon cancer cases, a male-to-female ratio that increases progressively from the proximal colon to the rectum,[1] and differences in age of onset by primary tumor site, with an earlier age of onset for rectal cancer[2] (supplementary figure 1 and supplementary table 1).

Assessing population stratification in the GWAS meta-analyses

To examine confounding due to residual population stratification, we inspected quantile-quantile (QQ) plots of tests statistics for the five GWAS meta-analyses (supplementary figure 2) and genomic control

inflation factors (λ_{GC}). λ_{GC} were 1.073, 1.077, 1.089, 1.090, and 1.101 for the GWAS meta-analyses for CRC case subgroups defined by proximal colon, distal colon, rectum, colon, and left-sided tumors subsites, respectively. The corresponding λ_{1000} values (λ_{GC} rescaled to a sample of 1,000 cases and 1,000 controls) were 1.007, 1.008, 1.007, 1.005, and 1.006. Given the highly polygenic genetic architecture of CRC, these estimates likely reflect the aggregate small effects of a large number of risk variants, although a small contribution of population stratification or bias cannot be excluded.

Genomic annotation of new GWAS loci and gene prioritization

To gain insight into the molecular mechanisms underlying the associations, and to nominate candidate causal variants and the most likely target gene(s) at each locus, we annotated all new risk 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 for non-coding and coding variants, (v) co-localization with expression quantitative trait loci (eQTL) signal.

We operationally defined a locus as the lead variant and variants in linkage disequilibrium ($r^2 \geq 0.8$) with the lead variant. We intersected variants with previously published functional and regulatory genomic annotations for normal colonic crypt epithelium and colonic mucosa tissue, and diverse CRC cell lines or tissue. Specifically, we assessed overlap with active enhancer regions identified by histone mark H3K27ac, and active regulatory regions identified by accessible chromatin identified through DNase I hypersensitive sites (DHSs) and ATAC-seq.[3,4] Additionally, to assess tissue-specificity we intersected with a compendium of ten groups of tissue-specific regulatory annotations for histone modifications (H3K4me1, H3K4me3, H3K9ac, and H3K27ac) obtained from various resources,[5,6] and previously derived by Finucane et al.[7] To assess gene-centric functional consequences of variants, we used the Ensembl Variant Effect Predictor tool to annotate variants relative to GENCODE and reference genome GRCh37.p13.[8] To determine whether variants potentially disrupt transcription factor (TF) binding, we

examined overlap with predicted transcription factor binding site (TFBS) motifs using the Haploreg v4.1 database,[9] and the chromatin immuno-precipitation sequence (ChIP-seq) binding sites for 161 transcription factors from the ENCODE Project.[10] We assessed colocalization with eQTL signals across tissues using GTEx V8 data.[11] To examine predicted functional impact, we annotated all locus variants with the CADD score (Phred scores >20 predicted as deleterious),[12] and we also annotated coding variants with the DANN score (scores >0.9 predicted as deleterious),[13] and PolyPhen-2 (benign, possibly damaging, probably damaging).[14] To evaluate expression in colorectal tissue, we interrogated the Human Protein Atlas[15] and GTEx V8 data.[11] Genes were prioritized based on biological relevance, expression in colorectal tissue, the presence of associated non-synonymous coding variants predicted to be deleterious, evidence from laboratory-based functional studies, somatic alterations, or familial syndromes linking them to CRC or cancer pathogenesis.

Data availability

All genotype data analyzed in this study have been previously published and have been deposited in the database of Genotypes and Phenotypes (dbGaP), which is hosted by NCBI, under accession numbers phs001415.v1.p1, phs001315.v1.p1, phs001078.v1.p1, and phs001903.v1.p1. The UK Biobank resource was accessed through application number 8614. CRC-relevant epigenome data were retrieved from the NCBI Gene Expression Omnibus (GEO) database under accession numbers GSE77737 and GSE36401.

Supplementary references

- 1 Murphy G, Devesa SS, Cross AJ, *et al.* Sex disparities in colorectal cancer incidence by anatomic subsite, race and age. *Int J Cancer* 2011;**128**:1668–75. doi:10.1002/ijc.25481
- 2 Murphy CC, Singal AG. Establishing a research agenda for early-onset colorectal cancer. *PLoS Med* 2018;**15**:e1002577. doi:10.1371/journal.pmed.1002577
- 3 Cohen AJ, Saiakhova A, Corradin O, *et al.* Hotspots of aberrant enhancer activity punctuate the colorectal cancer epigenome. *Nat Commun* 2017;**8**:14400. doi:10.1038/ncomms14400
- 4 Huyghe JR, Bien SA, Harrison TA, *et al.* Discovery of common and rare genetic risk variants for colorectal cancer. *Nat Genet* 2019;**51**:76–87. doi:10.1038/s41588-018-0286-6
- 5 Trynka G, Sandor C, Han B, *et al.* Chromatin marks identify critical cell types for fine mapping

- complex trait variants. *Nat Genet* 2013;**45**:124–30. doi:10.1038/ng.2504
- 6 Hnisz D, Abraham BJ, Lee TI, *et al*. Super-enhancers in the control of cell identity and disease. *Cell* 2013;**155**:934–47. doi:10.1016/j.cell.2013.09.053
- 7 Finucane HK, Bulik-Sullivan B, Gusev A, *et al*. Partitioning heritability by functional annotation using genome-wide association summary statistics. *Nat Genet* 2015;**47**:1228–35. doi:10.1038/ng.3404
- 8 McLaren W, Gil L, Hunt SE, *et al*. The ensembl variant effect predictor. *Genome Biol* 2016;**17**:122. doi:10.1186/s13059-016-0974-4
- 9 Ward LD, Kellis M. HaploReg: a resource for exploring chromatin states, conservation, and regulatory motif alterations within sets of genetically linked variants. *Nucleic Acids Res* 2012;**40**:D930–4. doi:10.1093/nar/gkr917
- 10 ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human genome. *Nature* 2012;**489**:57–74. doi:10.1038/nature11247
- 11 GTEx Consortium. The Genotype-Tissue Expression (GTEx) project. *Nat Genet* 2013;**45**:580–5. doi:10.1038/ng.2653
- 12 Kircher M, Witten DM, Jain P, *et al*. A general framework for estimating the relative pathogenicity of human genetic variants. *Nat Genet* 2014;**46**:310–5. doi:10.1038/ng.2892
- 13 Quang D, Chen Y, Xie X. DANN: a deep learning approach for annotating the pathogenicity of genetic variants. *Bioinformatics* 2015;**31**:761–3. doi:10.1093/bioinformatics/btu703
- 14 Adzhubei IA, Schmidt S, Peshkin L, *et al*. A method and server for predicting damaging missense mutations. *Nat Methods* 2010;**7**:248–9. doi:10.1038/nmeth0410-248
- 15 Uhlén M, Fagerberg L, Hallström BM, *et al*. Proteomics. Tissue-based map of the human proteome. *Science* 2015;**347**:1260419. doi:10.1126/science.1260419

FUNDING AND ACKNOWLEDGEMENTS

Funding:

Genetics and Epidemiology of Colorectal Cancer Consortium (GECCO): National Cancer Institute, National Institutes of Health, U.S. Department of Health and Human Services (U01 CA164930, U01 CA137088, R01 CA059045, R21 CA191312, R01 CA201407). Genotyping/Sequencing services were provided by the Center for Inherited Disease Research (CIDR) (X01-HG008596 and X01-HG007585). CIDR is fully funded through a federal contract from the National Institutes of Health to The Johns Hopkins University, contract number HHSN268201200008I. This research was funded in part through the NIH/NCI Cancer Center Support Grant P30 CA015704. PCS is supported by the National Institutes of Health (R01 CA160356, R01 CA193677, R01 CA204279 and R01 CA143237). This article is supported by the COST Action CA17118 and COST (European Cooperation in Science and Technology). COST is funded by the Horizon 2020 Framework Programme of the European Union.

ASTERISK: a Hospital Clinical Research Program (PHRC-BRD09/C) from the University Hospital Center of Nantes (CHU de Nantes) and supported by the Regional Council of Pays de la Loire, the Groupement des Entreprises Françaises dans la Lutte contre le Cancer (GEFLUC), the Association Anne de Bretagne Génétique and the Ligue Régionale Contre le Cancer (LRCC).

The ATBC Study is supported by the Intramural Research Program of the U.S. National Cancer Institute, National Institutes of Health, and by U.S. Public Health Service contract HHSN261201500005C from the National Cancer Institute, Department of Health and Human Services.

CLUE funding was from the National Cancer Institute (U01 CA86308, Early Detection Research Network; P30 CA006973), National Institute on Aging (U01 AG18033), and the American Institute for Cancer Research. The content of this publication does not necessarily reflect the views or policies of the Department of Health and Human Services, nor does mention of trade names, commercial products, or organizations imply endorsement by the US government.

COLO2&3: National Institutes of Health (R01 CA60987).

ColoCare: This work was supported by the National Institutes of Health (grant numbers R01 CA189184 (Li/Ulrich), U01 CA206110 (Ulrich/Li/Siegel/Figueiredo/Colditz), 2P30CA015704-40 (Gilliland), R01 CA207371 (Ulrich/Li)), the Matthias Lackas-Foundation, the German Consortium for Translational Cancer Research, and the EU TRANSCAN initiative.

The Colon Cancer Family Registry (CFR) Illumina GWAS was supported by funding from the National Cancer Institute, National Institutes of Health (grant numbers U01 CA122839, R01 CA143247 to G Casey). The Colon CFR/CORECT Affymetrix Axiom GWAS and OncoArray GWAS were supported by funding from National Cancer Institute, National Institutes of Health (grant number U19 CA148107 to S Gruber). The Colon CFR participant recruitment and collection of data and biospecimens used in this study were supported by the National Cancer Institute, National Institutes of Health (grant number U01 CA167551) and through cooperative agreements with the following Colon CFR centers: Australasian Colorectal Cancer Family Registry (NCI/NIH grant numbers U01 CA074778 and U01/U24 CA097735), USC Consortium Colorectal Cancer Family Registry (NCI/NIH grant numbers U01/U24 CA074799), Mayo Clinic Cooperative Family Registry for Colon Cancer Studies (NCI/NIH grant number U01/U24 CA074800), Ontario Familial Colorectal Cancer Registry (NCI/NIH grant number U01/U24 CA074783), Seattle Colorectal Cancer Family Registry (NCI/NIH grant number U01/U24 CA074794), and University of Hawaii Colorectal Cancer Family Registry (NCI/NIH grant number U01/U24 CA074806). Additional support for case ascertainment was provided from the Surveillance, Epidemiology and End Results (SEER) Program of the National Cancer Institute to Fred Hutchinson Cancer Research Center (Control Nos. N01-CN-67009 and N01-PC-35142, and Contract No. HHSN2612013000121), the Hawai'i Department of Health (Control Nos. N01-PC-67001 and N01-PC-35137, and Contract No. HHSN26120100037C, and the California Department of Public Health (contracts HHSN261201000035C awarded to the University of Southern California, and the following state cancer registries: AZ, CO, MN, NC, NH, and by the Victoria Cancer Registry and Ontario Cancer Registry.

COLON: The COLON study is sponsored by Wereld Kanker Onderzoek Fonds, including funds from grant 2014/1179 as part of the World Cancer Research Fund International Regular Grant Programme, by Alpe d'Huizes and the Dutch Cancer Society (UM 2012-5653, UW 2013-5927, UW2015-7946), and by TRANSCAN (JTC2012-MetaboCCC, JTC2013-FOCUS). The NQplus study is sponsored by a ZonMW investment grant (98-10030); by PREVIEW, the project PREvention of diabetes through lifestyle intervention and population studies in Europe and around the World (PREVIEW) project which received funding from the European Union Seventh Framework Programme (FP7/2007-2013) under grant no. 312057; by funds from TI Food and Nutrition (cardiovascular health theme), a public-private partnership on pre-competitive research in food and nutrition; and by FOOTBALL, the Food Biomarker Alliance, a project from JPI Healthy Diet for a Healthy Life.

Colorectal Cancer Transdisciplinary (CORECT) Study: The CORECT Study was supported by the National Cancer Institute, National Institutes of Health (NCI/NIH), U.S. Department of Health and

Human Services (grant numbers U19 CA148107, R01 CA81488, P30 CA014089, R01 CA197350, P01 CA196569, R01 CA201407, and R01 CA143237) and National Institutes of Environmental Health Sciences, National Institutes of Health (grant number T32 ES013678).

CORSA: This study was funded by FFG BRIDGE (grant 829675 to Andrea Gsur), the “Herzfelder’sche Familienstiftung” (grant to Andrea Gsur) and was supported by COST Action BM1206.

CPS-II: The American Cancer Society funds the creation, maintenance, and updating of the Cancer Prevention Study-II (CPS-II) cohort. This study was conducted with Institutional Review Board approval.

CRCGEN: Colorectal Cancer Genetics & Genomics, Spanish study was supported by Instituto de Salud Carlos III, co-funded by FEDER funds –a way to build Europe– (grants PI14-613 and PI09-1286), Agency for Management of University and Research Grants (AGAUR) of the Catalan Government (grant 2017SGR723), and Junta de Castilla y León (grant LE22A10-2). Sample collection of this work was supported by the Xarxa de Bancs de Tumors de Catalunya sponsored by Pla Director d’Oncologia de Catalunya (XBTC), Plataforma Biobancos PT13/0010/0013 and ICOBIOBANC, sponsored by the Catalan Institute of Oncology.

Czech Republic CCS: This project was supported by Grant Agency of the Ministry of Health of the Czech Republic (NV19-09-00237) and by Charles University Research Centre program UNCE/MED/006 “University Center of Clinical and Experimental Liver Surgery” and National Sustainability Program I (NPU I) Nr. LO1503 provided by the Ministry of Education Youth and Sports of the Czech Republic.

DACHS: This work was supported by the German Research Council (BR 1704/6-1, BR 1704/6-3, BR 1704/6-4, BR 1704/6-6, CH 117/1-1, HO 5117/2-1, HE 5998/2-1, KL 2354/3-1, RO 2270/8-1 and BR 1704/17-1), the Interdisciplinary Research Program of the National Center for Tumor Diseases (NCT), Germany, and the German Federal Ministry of Education and Research (01KH0404, 01ER0814, 01ER0815, 01ER1505A and 01ER1505B).

DALS: National Institutes of Health (R01 CA48998 to M. L. Slattery).

EDRN: This work is funded and supported by the NCI, EDRN Grant (U01 CA 84968-06).

EPIC: The coordination of EPIC is financially supported by the School of Public Health, Imperial College London and the International Agency for Research on Cancer. The national cohorts are supported by Danish Cancer Society (Denmark); Ligue Contre le Cancer, Institut Gustave Roussy, Mutuelle Générale de l’Education Nationale, Institut National de la Santé et de la Recherche Médicale (INSERM) (France); German Cancer Aid, German Cancer Research Center (DKFZ), Federal Ministry of Education and Research (BMBF), Deutsche Krebshilfe, Deutsches Krebsforschungszentrum and Federal Ministry of Education and Research (Germany); the Hellenic Health Foundation (Greece); Associazione Italiana per la Ricerca sul Cancro-AIRC-Italy and National Research Council (Italy); Dutch Ministry of Public Health, Welfare and Sports (VWS), Netherlands Cancer Registry (NKR), LK Research Funds, Dutch Prevention Funds, Dutch ZON (Zorg Onderzoek Nederland), World Cancer Research Fund (WCRF), Statistics Netherlands (The Netherlands); ERC-2009-AdG 232997 and Nordforsk, Nordic Centre of Excellence programme on Food, Nutrition and Health (Norway); Health Research Fund (FIS), PI13/00061 to Granada, PI13/01162 to EPIC-Murcia, Regional Governments of Andalucía, Asturias, Basque Country, Murcia and Navarra, ISCIII RETIC (RD06/0020) (Spain); Swedish Cancer Society, Swedish Research Council and County Councils of Skåne and Västerbotten (Sweden); Cancer Research UK (14136 to EPIC-Norfolk; C570/A16491 and C8221/A19170 to EPIC-Oxford), Medical Research Council (1000143 to EPIC-Norfolk, MR/M012190/1 to EPIC-Oxford) (United Kingdom).

EPICOLON: This work was supported by grants from Fondo de Investigación Sanitaria/FEDER (PI08/0024, PI08/1276, PS09/02368, P111/00219, P111/00681, PI14/00173, PI14/00230, PI17/00509, 17/00878, Acción Transversal de Cáncer), Xunta de Galicia (PGIDIT07PXIB9101209PR), Ministerio de Economía y Competitividad (SAF07-64873, SAF 2010-19273, SAF2014-54453R), Fundación Científica de la Asociación Española contra el Cáncer (GCB13131592CAST), Beca Grupo de Trabajo “Oncología” AEG (Asociación Española de Gastroenterología), Fundación Privada Olga Torres, FP7 CHIBCHA Consortium, Agència de Gestió d’Ajuts Universitaris i de Recerca (AGAUR, Generalitat de Catalunya, 2014SGR135, 2014SGR255, 2017SGR21, 2017SGR653), Catalan Tumour Bank Network (Pla Director d’Oncologia, Generalitat de Catalunya), PERIS (SLT002/16/00398, Generalitat de Catalunya), CERCA Programme (Generalitat de Catalunya) and COST Action CA17118, supported by COST (European Cooperation in Science and Technology). CIBERehd is funded by the Instituto de Salud Carlos III.

ESTHER/VERDI. This work was supported by grants from the Baden-Württemberg Ministry of Science, Research and Arts and the German Cancer Aid.

Harvard cohorts (HPFS, NHS, PHS): HPFS is supported by the National Institutes of Health (P01 CA055075, UM1 CA167552, U01 CA167552, R01 CA137178, R01 CA151993, R35 CA197735, K07 CA190673, and P50 CA127003), NHS by the National Institutes of Health (R01 CA137178, P01 CA087969, UM1 CA186107, R01 CA151993, R35 CA197735, K07 CA190673, and P50 CA127003) and PHS by the National Institutes of Health (R01 CA042182).

Hawaii Adenoma Study: NCI grants R01 CA72520.

Kentucky: This work was supported by the following grant support: Clinical Investigator Award from Damon Runyon Cancer Research Foundation (CI-8); NCI R01CA136726.

LCCS: The Leeds Colorectal Cancer Study was funded by the Food Standards Agency and Cancer Research UK Programme Award (C588/A19167).

MCCS cohort recruitment was funded by VicHealth and Cancer Council Victoria. The MCCS was further supported by Australian NHMRC grants 209057, 396414 and 1074383, and by infrastructure provided by Cancer Council Victoria. Cases and their vital status were ascertained through the Victorian Cancer Registry (VCR) and the Australian Institute of Health and Welfare (AIHW), including the National Death Index and the Australian Cancer Database.

MEC: National Institutes of Health (R37 CA54281, P01 CA033619, R01 CA063464, and U01 CA164973).

MECC: This work was supported by the National Institutes of Health, U.S. Department of Health and Human Services (R01 CA81488 to SBG, R01 CA197350 to SBG, U19 CA148107 to SBG, N01 CN043302 assigned to SBG, 5P01 CA196569 to WG, P30 CA014089 to SBG, R01 CA144040 to SDM, and P50 CA150964 to SDM), as well as funding from the Ravitz Foundation, the Irving Weinstein Foundation, the Anton B. Burg Foundation, the Jane and Kris Popovich Chair in Cancer Research, and a generous gift from Daniel and Maryann Fong.

MSKCC: The work at Memorial Sloan Kettering Cancer Center in New York was supported by the Robert and Kate Niehaus Center for Inherited Cancer Genomics and the Romeo Milio Foundation. It is also supported by the National Cancer Institute-designated Comprehensive Cancer Center (grant number P30 CA008748).

NCCCS I & II: We acknowledge funding support for this project from the National Institutes of Health, R01 CA66635 and P30 DK034987.

NFCCR: This work was supported by an Interdisciplinary Health Research Team award from the Canadian Institutes of Health Research (CRT 43821); the National Cancer Institute, US National Institutes of Health (grant number U01 CA74783 to S Gallinger); and National Cancer Institute of Canada grants (18223 and 18226). Funding was provided to Michael O. Woods by the Canadian Cancer Society Research Institute.

NSHDS: This work was funded by the Swedish Research Council, Swedish Cancer Society; Cancer Research Foundation and Lion's Cancer Research Foundation in Northern Sweden; J C Kempe Memorial Fund; Faculty of Medicine, Umeå University, Umeå, Sweden; and the Västerbotten County Council, Sweden.

OFCCR: National Institutes of Health, through funding allocated to the Ontario Registry for Studies of Familial Colorectal Cancer (U01 CA74783 to S Gallinger); see CCFR section above. Additional funding toward genetic analyses of OFCCR includes the Ontario Research Fund, the Canadian Institutes of Health Research, and the Ontario Institute for Cancer Research, through generous support from the Ontario Ministry of Research and Innovation.

OSUWMC: OCCPI funding was provided by Pelotonia and HNPCC funding was provided by the NCI (CA16058 and CA67941).

PLCO: Intramural Research Program of the Division of Cancer Epidemiology and Genetics and supported by contracts from the Division of Cancer Prevention, National Cancer Institute, NIH, DHHS. Funding was provided by National Institutes of Health (NIH), Genes, Environment and Health Initiative (GEI) Z01 CP 010200, NIH U01 HG004446, and NIH GEI U01 HG 004438.

PMH-SCCFR: National Cancer Institute, National Institutes of Health (grant numbers U01 CA074794 to J Potter and R01 CA076366 to P.A. Newcomb).

SEARCH: The University of Cambridge has received salary support in respect of PDPP from the NHS in the East of England through the Clinical Academic Reserve. Cancer Research UK (C490/A16561); the UK National Institute for Health Research Biomedical Research Centres at the University of Cambridge.

SELECT: The Selenium and Vitamin E Cancer Prevention Trial (SELECT) was supported by the National Cancer Institute of the National Institutes of Health under Award Numbers UM1CA182883 and U10CA37429.

SMS: This work was supported by the National Cancer Institute (grant P01 CA074184 to J.D.P. and P.A.N., grants R01 CA097325, R03 CA153323, and K05 CA152715 to P.A.N., and the National Center for Advancing Translational Sciences at the National Institutes of Health (grant KL2 TR000421 to A.N.B.-H.)

The Swedish Low-risk Colorectal Cancer Study: The study was supported by grants from the Swedish research council; K2015-55X-22674-01-4, K2008-55X-20157-03-3, K2006-72X-20157-01-2 and the Stockholm County Council (ALF project).

Swedish Mammography Cohort and Cohort of Swedish Men: This work is supported by grants from the Swedish Cancer Foundation, the Swedish Research Council for the Swedish Infrastructure for Medical

Population-based Life-course Environmental Research (SIMPLER) and Karolinska Institute's Distinguished Professor Award to Alicja Wolk.

VITAL: National Institutes of Health (K05 CA154337).

WHI: The WHI program is funded by the National Heart, Lung, and Blood Institute, National Institutes of Health, U.S. Department of Health and Human Services through contracts HHSN268201100046C, HHSN268201100001C, HHSN268201100002C, HHSN268201100003C, HHSN268201100004C, and HHSN271201100004C.

Acknowledgements:

ASTERISK: We are very grateful to Dr. Bruno Buecher without whom this project would not have existed. We also thank all those who agreed to participate in this study, including the patients and the healthy control persons, as well as all the physicians, technicians and students.

CLUE: We appreciate the continued efforts of the staff members at the Johns Hopkins George W. Comstock Center for Public Health Research and Prevention in the conduct of the CLUE II study. We thank the participants in CLUE. Cancer incidence data for CLUE were provided by the Maryland Cancer Registry, Center for Cancer Surveillance and Control, Maryland Department of Health, 201 W. Preston Street, Room 400, Baltimore, MD 21201, <http://phpa.dhmdh.maryland.gov/cancer>, 410-767-4055. We acknowledge the State of Maryland, the Maryland Cigarette Restitution Fund, and the National Program of Cancer Registries of the Centers for Disease Control and Prevention for the funds that support the collection and availability of the cancer registry data.

COLON and NQplus: the authors would like to thank the COLON and NQplus investigators at Wageningen University & Research and the involved clinicians in the participating hospitals.

CORSA: We kindly thank all those who contributed to the screening project Burgenland against CRC. Furthermore, we also thank all individuals who agreed to participate in the CORSA study, all physicians and students.

CPS-II: The authors thank the CPS-II participants and Study Management Group for their invaluable contributions to this research. The authors would also like to acknowledge the contribution to this study from central cancer registries supported through the Centers for Disease Control and Prevention National Program of Cancer Registries, and cancer registries supported by the National Cancer Institute Surveillance Epidemiology and End Results program.

Czech Republic CCS: We are thankful to all clinicians in major hospitals in the Czech Republic, without whom the study would not be practicable. We are also sincerely grateful to all patients participating in this study.

DACHS: We thank all participants and cooperating clinicians, and Ute Handte-Daub, Ansgar Brandhorst, Utz Benseid, Muhabbet Celik and Ursula Eilber for excellent technical assistance.

EDRN: We acknowledge all the following contributors to the development of the resource: University of Pittsburgh School of Medicine, Department of Gastroenterology, Hepatology and Nutrition: Lynda Dzubinski; University of Pittsburgh School of Medicine, Department of Pathology: Michelle Bisceglia; and University of Pittsburgh School of Medicine, Department of Biomedical Informatics.

EPICOLON: We are sincerely grateful to all patients, the Spanish National DNA Bank (BNADN) and Biobank of Hospital Clínic-IDIBAPS.

Harvard cohorts (HPFS, NHS, PHS): The study protocol was approved by the institutional review boards of the Brigham and Women's Hospital and Harvard T.H. Chan School of Public Health, and those of participating registries as required. We would like to thank the participants and staff of the HPFS, NHS and PHS for their valuable contributions as well as the following state cancer registries for their help: AL, AZ, AR, CA, CO, CT, DE, FL, GA, ID, IL, IN, IA, KY, LA, ME, MD, MA, MI, NE, NH, NJ, NY, NC, ND, OH, OK, OR, PA, RI, SC, TN, TX, VA, WA, WY. The authors assume full responsibility for analyses and interpretation of these data.

Kentucky: We would like to acknowledge the staff at the Kentucky Cancer Registry.

LCCS: We acknowledge the contributions of Jennifer Barrett, Robin Waxman, Gillian Smith and Emma Northwood in conducting this study.

NCCCS I & II: We would like to thank the study participants, and the NC Colorectal Cancer Study staff.

NFCCR: The authors wish to acknowledge the contribution of Alexandre Belisle and the genotyping team of the McGill University and Génome Québec Innovation Centre, Montréal, Canada, for genotyping the Sequenom panel in the NFCCR samples.

NSHDS: We thank all participants in the NSHDS cohorts and the staff at the Department of Biobank Research, Umeå University, as well as Kerstin Näslund, formerly of the Department of Medical Biosciences, and Inger Cullman, Department of Chemistry, both at Umeå University for excellent technical assistance.

PLCO: The authors thank the PLCO Cancer Screening Trial screening center investigators and the staff from Information Management Services Inc and Westat Inc. Most importantly, we thank the study participants for their contributions that made this study possible.

PMH: The authors would like to thank the study participants and staff of the Hormones and Colon Cancer study.

SEARCH: We thank the SEARCH team.

SELECT: We thank the research and clinical staff at the sites that participated on SELECT study, without whom the trial would not have been successful. We are also grateful to the 35,533 dedicated men who participated in SELECT.

WHI: The authors thank the WHI investigators and staff for their dedication, and the study participants for making the program possible. A full listing of WHI investigators can be found at: <http://www.whi.org/researchers/Documents%20%20Write%20a%20Paper/WHI%20Investigator%20Short%20List.pdf>