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], David Duggan[11], Sophia Harlid[12], Liher Imaz[13], Hyun Min Kang[7], David M Levine[14], Vittorio Perduca[15,16], Aurora Perez-Cornago[17], Lori C Sakoda[1,18], Fredrick R Schumacher[19], Martha L Slattery[20], Amanda E Toland[21], Franzel JB van Duijnhoven[22], Bethany Van Guelpen[12], Volker Arndt[23], Antonio Agudo[24], Demetrius Albanes[25], M Henar Alonso[26–28], Kristin Anderson[29], Coral Arnau-Collell[30], Barbara Banbury[1], Michael C Bassik[31], Sonja I Berndt[25], Stéphane Bézieau[32], D Timothy Bishop[33], Juergen Boehm[34], Heiner Boeing[35], Marie-Christine Boutron-Ruault[36,37], Hermann Brenner[23,38,39], Stefanie Brezina[40], Stephan Buch[41], Daniel D Buchanan[42–44], Andrea Burnett-Hartman[45], Bette J Caan[46], Peter T Campbell[47], Prudence Carr[48], Antoni Castells[30], Sergi Castellví-Bel[30], Andrew T Chan[49–54], 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 JM Feskens[22], 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], John S Grove[63], Andrea Gsur[40], Marc J Gunter[64], Robert W Haile[3], Jochen Hampe[41], Michael Hoffmeister[23], John L Hopper[42,65], Wan-Ling Hsu[14], Wen-Yi Huang[25], Thomas J Hudson[66], Mazda Jenab[67], Mark A Jenkins[42], Amit D Joshi[51,53], Temitope O Keku[68], Charles Kooperberg[1], Tilman Kuhn[55], Sébastien Küry[32], Loic Le Marchand[63], Flavio Lejbkowicz[69–71], 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[64], 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[34], Kala Visvanathan[87], Pavel Vodicka[99–101], Ludmila Vodickova[99–101], Veronika Vymetalkova[99–101], Korbinian Weigl[23,39,102], 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[31,107], Graham Casey[108], Stephen B Gruber[6], Li Hsu[1,14], Victor Moreno[26–28], Richard B Hayes[109], Polly A Newcomb[1,86], Ulrike Peters[1,86]

1. Public Health Sciences Division, Fred Hutchinson Cancer Research Center, Seattle, Washington, USA.
2. Division of Human Genetics, Department of Internal Medicine, The Ohio State University Comprehensive Cancer Center, Columbus, Ohio, USA.
3. Department of Medicine, Samuel Oschin Comprehensive Cancer Institute, Cedars-Sinai Medical Center, Los Angeles, California, USA.
4. Department of Preventive Medicine, Keck School of Medicine, University of Southern California,

Los Angeles, California, USA.

5. Department of Cancer Epidemiology, H. Lee Moffitt Cancer Center and Research Institute, Tampa, Florida, USA.
6. Department of Preventive Medicine, USC Norris Comprehensive Cancer Center, Keck School of Medicine, University of Southern California, Los Angeles, California, USA.
7. Department of Biostatistics and Center for Statistical Genetics, University of Michigan, Ann Arbor, Michigan, USA.
8. Department of Medicine, University of North Carolina School of Medicine, Chapel Hill, North Carolina, USA.
9. Department of Epidemiology and Biostatistics, Imperial College London, London, UK.
10. Department of Human Genetics, Graduate School of Public Health, University of Pittsburgh, and UPMC Hillman Cancer Center, Pittsburgh, PA.
11. Translational Genomics Research Institute - An Affiliate of City of Hope, Phoenix, Arizona, USA.
12. Department of Radiation Sciences, Oncology Unit, Umeå University, Umeå, Sweden.
13. Public Health Division of Gipuzkoa, Health Department of Basque Country, Spain.
14. Department of Biostatistics, University of Washington, Seattle, Washington, USA.
15. Laboratoire de Mathématiques Appliquées MAP5 (UMR CNRS 8145), Université Paris Descartes, Paris, France.
16. CESP (Inserm U1018), Facultés de Médicine Université Paris-Sud, UVSQ, Université Paris-Saclay, Gustave Roussy, 94805, Villejuif, France.
17. Cancer Epidemiology Unit, Nuffield Department of Population Health, University of Oxford, Oxford, UK.
18. Division of Research, Kaiser Permanente Northern California, Oakland, California, USA.
19. Department of Population and Quantitative Health Sciences, Case Western Reserve University, Cleveland, Ohio, USA.
20. Department of Internal Medicine, University of Utah, Salt Lake City, Utah, USA.

21. Departments of Cancer Biology and Genetics and Internal Medicine, Comprehensive Cancer Center, The Ohio State University, Columbus, Ohio, USA.

22. Division of Human Nutrition and Health, Wageningen University & Research, Wageningen, The Netherlands.

23. Division of Clinical Epidemiology and Aging Research, German Cancer Research Center (DKFZ), Heidelberg, Germany.

24. Unit of Nutrition and Cancer, Cancer Epidemiology Research Program, Catalan Institute of Oncology-IDIBELL, L'Hospitalet de Llobregat, Barcelona, Spain.

25. Division of Cancer Epidemiology and Genetics, National Cancer Institute, National Institutes of Health, Bethesda, Maryland, USA.

26. Cancer Prevention and Control Program, Catalan Institute of Oncology-IDIBELL, L'Hospitalet de Llobregat, Barcelona, Spain.

27. CIBER de Epidemiología y Salud Pública (CIBERESP), Madrid, Spain.

28. Department of Clinical Sciences, Faculty of Medicine, University of Barcelona, Barcelona, Spain.

29. Division of Epidemiology and Community Health, University of Minnesota, Minneapolis, Minnesota, USA.

30. Gastroenterology 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.

31. Department of Genetics, Stanford University, Stanford, California, USA.

32. Service de Génétique Médicale, Centre Hospitalier Universitaire (CHU) Nantes, Nantes, France.

33. Leeds Institute of Medical Research at St James's, University of Leeds, Leeds, UK.

34. Huntsman Cancer Institute and Department of Population Health Sciences, University of Utah, Salt Lake City, Utah, USA.

35. Department of Epidemiology, German Institute of Human Nutrition (DIfE), Potsdam-Rehbrücke, Germany.

36. Inserm U1018, Center for Research in Epidemiology and Population Health (CESP), Gustave Roussy, Villejuif, France.

37. Paris-South Saclay University, Villejuif, France.

38. Division of Preventive Oncology, German Cancer Research Center (DKFZ) and National Center for Tumor Diseases (NCT), Heidelberg, Germany.

39. German Cancer Consortium (DKTK), German Cancer Research Center (DKFZ), Heidelberg, Germany.

40. Institute of Cancer Research, Department of Medicine I, Medical University of Vienna, Vienna, Austria.

41. Department of Medicine I, University Hospital Dresden, Technische Universität Dresden (TU Dresden), Dresden, Germany.

42. Centre for Epidemiology and Biostatistics, Melbourne School of Population and Global Health, The University of Melbourne, Melbourne, Victoria, Australia.

43. Colorectal Oncogenomics Group, Department of Clinical Pathology, The University of Melbourne, Parkville, Victoria, Australia.

44. Genomic Medicine and Family Cancer Clinic, Royal Melbourne Hospital, Parkville, Victoria, Australia.

45. Institute for Health Research, Kaiser Permanente Colorado, Denver, Colorado, USA.

46. Division of Research, Kaiser Permanente Medical Care Program, Oakland, California, USA.

47. Behavioral and Epidemiology Research Group, American Cancer Society, Atlanta, Georgia, USA.

48. Division of Clinical Epidemiology, German Cancer Research Center (DKFZ), Heidelberg, Germany.

49. Division of Gastroenterology, Massachusetts General Hospital and Harvard Medical School, Boston, Massachusetts, USA.

50. Channing Division of Network Medicine, Brigham and Women's Hospital and Harvard Medical School, Boston, Massachusetts, USA.

51. Clinical and Translational Epidemiology Unit, Massachusetts General Hospital and Harvard Medical School, Boston, Massachusetts, USA.

52. Broad Institute of Harvard and MIT, Cambridge, Massachusetts, USA.

53. Department of Epidemiology, Harvard T.H. Chan School of Public Health, Harvard University, Boston, Massachusetts, USA.

54. Department of Immunology and Infectious Diseases, Harvard T.H. Chan School of Public Health, Harvard University, Boston, Massachusetts, USA.

55. Division of Cancer Epidemiology, German Cancer Research Center (DKFZ), Heidelberg, Germany.

56. Cancer Epidemiology Group, University Medical Centre Hamburg-Eppendorf, University Cancer Centre Hamburg (UCCH), Hamburg, Germany.

57. Department of Cancer Biology and Genetics and the Comprehensive Cancer Center, The Ohio State University, Columbus, Ohio, USA.

58. Department of Public Health and Primary Care, University of Cambridge, Cambridge, UK.

59. Cancer Epidemiology Division, Cancer Council Victoria, Melbourne, Victoria, Australia.

60. Lunenfeld Tanenbaum Research Institute, Mount Sinai Hospital, University of Toronto, Toronto, Ontario, Canada.

61. SWOG Statistical Center, Fred Hutchinson Cancer Research Center, Seattle, Washington, USA.

62. Clinical Research Division, Fred Hutchinson Cancer Research Center, Seattle, Washington, USA.

63. University of Hawaii Cancer Research Center, Honolulu, Hawaii, USA.

64. Nutrition and Metabolism Section, International Agency for Research on Cancer, World Health Organization, Lyon, France.

65. Department of Epidemiology, School of Public Health and Institute of Health and Environment, Seoul National University, Seoul, South Korea.

66. Ontario Institute for Cancer Research, Toronto, Ontario, Canada.

67. International Agency for Research on Cancer, World Health Organization, Lyon, France.

68. Center for Gastrointestinal Biology and Disease, University of North Carolina, Chapel Hill, North Carolina, USA.

69. The Clalit Health Services, Personalized Genomic Service, Carmel, Haifa, Israel.

70. Department of Community Medicine and Epidemiology, Lady Davis Carmel Medical Center, Haifa, Israel.

71. Clalit National Cancer Control Center, Haifa, Israel.

72. Department of Family Medicine, University of Virginia, Charlottesville, Virginia, USA.

73. Institute of Epidemiology, PopGen Biobank, Christian-Albrechts-University Kiel, Kiel, Germany.

74. Department of Clinical Genetics, Karolinska University Hospital, Stockholm, Sweden.

75. Department of Molecular Medicine and Surgery, Karolinska Institutet, Stockholm, Sweden.

76. Department of Health Science Research, Mayo Clinic, Scottsdale, Arizona, USA.

77. Department of Public Health Solutions, National Institute for Health and Welfare, Helsinki, Finland.

78. Departments of Medicine and Genetics, Case Comprehensive Cancer Center, Case Western Reserve University, and University Hospitals of Cleveland, Cleveland, Ohio, USA.

79. Department of Pathology, School of Medicine, Umm Al-Qura'a University, Saudi Arabia.

80. Clinical Genetics Service, Department of Medicine, Memorial Sloan Kettering Cancer Center, New York, New York, USA.

81. Department of Medicine, Weill Cornell Medical College, New York, New York, USA.

82. Program in MPE Molecular Pathological Epidemiology, Department of Pathology, Brigham and Women's Hospital, Harvard Medical School, Boston, Massachusetts, USA.

83. Department of Oncologic Pathology, Dana-Farber Cancer Institute, Boston, Massachusetts, USA.

84. Dipartimento di Medicina Clinica e Chirurgia, Federico II University, Naples, Italy.

85. Clinical Epidemiology Unit, Faculty of Medicine, Memorial University, St. John's, Newfoundland, Canada.

86. Department of Epidemiology, University of Washington, Seattle, Washington, USA.

87. Department of Epidemiology, Johns Hopkins Bloomberg School of Public Health, Johns Hopkins University, Baltimore, Maryland, USA.

88. Department of Public Health Sciences, School of Medicine, University of California Davis, Davis, California, USA.

89. Division of Epidemiology, Vanderbilt Epidemiology Center, Vanderbilt University School of Medicine, Nashville, Tennessee, USA.

90. Ruth and Bruce Rappaport Faculty of Medicine, Technion-Israel Institute of Technology, Haifa, Israel.

91. School of Public Health, Imperial College London, London, UK.

92. Department of General Surgery, University Hospital Rostock, Rostock, Germany.

93. Department of Medicine and Epidemiology, University of Pittsburgh Medical Center, Pittsburgh, Pennsylvania, USA.

94. Division of Cancer Control and Population Sciences, National Cancer Institute, Bethesda, Maryland, USA.

95. Department of Nutrition, Harvard T.H. Chan School of Public Health, Harvard University, Boston, Massachusetts, USA.

96. Division of Laboratory Genetics, Department of Laboratory Medicine and Pathology, Mayo Clinic, Rochester, Minnesota, USA.

97. Hellenic Health Foundation, Athens, Greece.

98. WHO 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, Greece.

99. Department of Molecular Biology of Cancer, Institute of Experimental Medicine of the Czech Academy of Sciences, Prague, Czech Republic.

100. Institute of Biology and Medical Genetics, First Faculty of Medicine, Charles University, Prague, Czech Republic.

101. Faculty of Medicine and Biomedical Center in Pilsen, Charles University, Pilsen, Czech Republic.

102. Medical Faculty, University of Heidelberg, Germany.

103. Institute of Environmental Medicine, Karolinska Institutet, Stockholm, Sweden.

104. Memorial University of Newfoundland, Discipline of Genetics, St. John's, Canada.

105. Department of Genome Sciences, University of Washington, Seattle, Washington, USA.

106. Department of Genetics and Genome Sciences, Case Western Reserve University School of Medicine, Case Comprehensive Cancer Center, Cleveland, Ohio, USA.

107. Department of Computer Science, Stanford University, Stanford, California, USA.

108. Center for Public Health Genomics, University of Virginia, Charlottesville, Virginia, USA.

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

Correspondence to:

Ulrike Peters, PhD, MPH, Fred Hutchinson Cancer Research Center, 1100 Fairview Avenue North, M4-B402, PO Box 19024, Seattle, Washington 98109. e-mail: upeters@fredhutch.org; fax: (206) 667-7850.

Disclosures:

The authors disclose no conflicts of interests.

Disclaimer:

Where authors are identified as personnel of the International Agency for Research on Cancer / World Health Organization, 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 / World Health Organization.

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 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 sporadic 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 characterized effect heterogeneity at CRC risk loci using multinomial modeling.

Results We identified 13 loci that reached genome-wide significance ($P<5\times10^{-8}$) and that were not reported by previous GWAS 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 tumor.

Keywords: colorectal cancer, genetic risk factors, genetic heterogeneity, proximal colon, distal colorectum

Significance of this study

What is already known about this subject?

- Heterogeneity among colorectal cancer (CRC) tumors 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 (GWAS) 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 tumor is lacking.

What are the new findings?

- In this large consortium-based study, we analyzed 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 etiologies.
- We discovered 13 new loci at genome-wide significance ($P<5\times10^{-8}$) that were specific to certain anatomical sublocations and that were not reported by previous GWAS 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 the genetic architectures of proximal and distal CRC are partly distinct.
- Taken together, our results further support the idea that tumors arising in different anatomical sublocations of the colorectum may have distinct etiologies.

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

- Our results provide an informative resource for understanding the differential role that genes and pathways may play in the mechanisms of proximal and distal CRC carcinogenesis.

- The new insights into the etiologies of proximal and distal CRC may inform the development of new precision prevention strategies, including individualized screening recommendations and the discovery of novel drug targets and repurposable drug candidates for chemoprevention.
- Our findings suggest that future studies of etiological risk factors for CRC and molecular mechanisms of carcinogenesis should take into consideration the anatomical sublocation of the colorectal tumor.

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 United States alone.[1] CRCs that arise proximal (right) or distal (left) to the splenic flexure differ in age- and sex-specific incidence rates, clinical, pathological and tumor 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 etiologic 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 sporadic 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 genetic association signals for overall sporadic CRC risk have been identified, over half of which were only identified in the past few years.[7–10] Three decades ago, based on observed similarities between Lynch syndrome and proximal sporadic CRC, and between Familial Adenomatous Polyposis (FAP) and distal sporadic CRC, Bufill proposed the existence of two distinct genetic categories of sporadic CRC according to the location of the primary tumor.[2] However, given that genetic variants that influence sporadic CRC

risk typically have small effect sizes, until very recently, sample sizes of GWASs did not provide adequate statistical power to conduct meaningful subsite analyses. As a consequence, discovery GWASs to detect genetic variants that are specific for CRC case subgroups defined by anatomic subsite of the primary tumor have not been reported yet. Similarly, a comprehensive analysis of the extent to which allelic risk of the known GWAS-identified genetic variants differs by the anatomic subsite of the primary tumor is lacking.

To address the major gap in our knowledge of the differential role that genetic variants, genes and pathways play in the mechanisms of proximal and distal CRC carcinogenesis, we conducted a large consortium-based study that included 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 CRC case subgroups defined by the location of their primary tumor within the colorectum. Next, we systematically characterized heterogeneity of allelic effects between primary tumor 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 the online supplementary materials.

Samples and genotypes

This study included clinical and genotype data for 48,214 CRC cases and 64,159 controls from three consortia: the Genetics and Epidemiology of Colorectal Cancer Consortium (GECCO), the Colorectal Cancer Transdisciplinary Study (CORECT), and Colorectal Cancer Family Registry (CCFR). Supplementary table 1 provides details on sample numbers and demographic characteristics by study. All analyses were restricted to genetically inferred European-ancestry participants. Across studies, participant recruitment occurred between the early 1990s and the 2010s. Details of all 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 tumor anatomic sublocation definitions

We defined proximal colon cancer as any primary tumor arising in the cecum, ascending colon, hepatic flexure, or transverse colon; distal colon cancer as any primary tumor arising in the splenic flexure, descending colon, or sigmoid colon; and rectal cancer as any primary tumor arising in the rectum or rectosigmoid junction. 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 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. Supplementary figure 1 and supplementary table 1 summarize distributions of age of diagnosis by sex and primary tumor site.

Statistical analysis

GWAS meta-analyses

We imputed all genotype data sets to the Haplotype Reference Consortium (HRC) panel, which by combining sequencing data from 32,488 individuals, enables accurate imputation of single nucleotide variants (SNVs) with minor allele frequencies (MAF) as low as 0.1%.[13] In brief, we phased all genotyping array data sets using SHAPEIT2[14] and used the Michigan Imputation Server[15] to impute to the HRC panel. Within each data set, variants with an imputation accuracy $r^2 \geq 0.3$ and minor allele count (MAC) ≥ 50 were tested for association with CRC case subgroup. Variants that only passed filters in a single genotype data set were excluded. We assumed an additive genetic model using the imputed

genotype dosage in a logistic regression model adjusted for age, sex, and study or genotyping project-specific covariates, including principal components to adjust for population structure. Details of the covariate corrections for each data set have been published previously.[8] Because Wald tests can be anti-conservative 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 odds ratio (OR) estimates and 95% confidence intervals (CIs) are based on an inverse variance-weighted fixed-effects meta-analysis.

Heterogeneity in allelic effect sizes between tumor anatomic sublocations

To characterize tumor subsite-specificity and effect size heterogeneity across tumor subsites for newly identified loci, as well as for established loci associated with overall CRC risk in previous GWAS meta-analyses, we examined the association evidence in three different ways. First, for each index variant we created forest plots of OR estimates with 95% CIs from the GWAS meta-analyses for proximal colon, distal colon, and rectal cancer. Second, we tested for heterogeneity using multinomial logistic regression analysis. In brief, after pooling of data sets, we performed a likelihood ratio test comparing a model in which the ORs for the risk variant were allowed to vary across tumor subsites, to a model in which the ORs were constrained to be the same across tumor sites. Third, we used a multinomial logistic regression-based model selection approach to assess which configuration of tumor 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 tumor subsites (supplementary 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.[17] For these variants, we do not report a single best model.

Analyses were carried out using the VGAM R package.[18] The list of index variants for previously published CRC risk signals is based on Huyghe *et al*.[8].

Genomic annotation of new GWAS loci and gene prioritization

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 eQTL signals. Genes were further prioritized based on biological relevance, colorectal tissue expression, 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. Detailed methods and references of the databases queried are provided in the online supplementary 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 CRC case subgroups defined by the location of their primary tumor 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 characterized heterogeneity of allelic effects between primary tumor 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 SNVs were analyzed. Inspection of genomic control inflation factors (λ_{GC} and λ_{1000}) and quantile-quantile (QQ) plots of test statistics indicated no residual population stratification issues (online supplementary materials and supplementary figure 2). Across CRC tumor subsites, we identified 13 CRC risk loci that mapped outside of regions previously implicated by GWASs for overall CRC risk (closest known locus 3.1 megabases away) and that passed the genome-wide significance threshold of $P < 5\times10^{-8}$ in at least one of the meta-analyses (table 1; figure 1; supplementary figures 3 and 4). Seven of these 13 new loci passed a Bonferroni-adjusted genome-wide significance threshold correcting for the five case subgroups analyzed (table 1). All lead variants were well imputed (minimum average imputation r^2 of 0.788), had MAF >1%, and displayed no significant heterogeneity between genotyping sample sets (Cochran's Q test for heterogeneity $P > 0.05$; table 1).

The novel associations showing the strongest statistical evidence were obtained for proximal colon cancer and mapped near *MLH1* on 3p22.2 (rs1800734, $P=3.8\times10^{-18}$) and near *BCL11B* on 14q32.2 (rs80158569, $P=8.6\times10^{-11}$). These loci showed strongly proximal cancer-specific associations. The proximal colon analysis yielded an additional locus at 14q32.12 (rs61975764, $P=2.8\times10^{-8}$) that showed attenuated effects for cancers at other sites of the colorectum (figure 1 and supplementary table 3). Most new loci (six) were discovered in the left-sided CRC analysis: 2q21.3 (rs1446585, $P=3.3\times10^{-8}$), near *CDX1* on 5q32 (rs2302274, $P=4.9\times10^{-9}$), near *KLF14* on 7q32.3 (rs73161913, $P=1.3\times10^{-9}$), 10q23.31 (rs7071258. $P=8.4\times10^{-9}$), 19p13.3 (rs62131228, $P=2.4\times10^{-8}$), and near *BMP7* on 20q13.31 (rs6014965, $P=4.5\times10^{-9}$). The rectal cancer analysis identified an additional locus near *PYGL* on 14q22.1 (rs28611105, $P=4.7\times10^{-9}$) that showed an attenuated effect for distal colon cancer (figure 1 and supplementary table 3). No additional new loci were detected in the distal colon analysis. The colon cancer analysis identified three new genome-wide significant loci for colon cancer near *PTGER3* on 1p31.1 (rs3124454, $P=1.4\times10^{-8}$), 3p21.2 (rs353548, $P=1.3\times10^{-8}$), and 22q13.31 (rs736037, $P=2.8\times10^{-8}$).

Table 1. New genome-wide significant colorectal cancer risk loci identified by genome-wide association analysis of case subgroups defined by primary tumor anatomic subsite.

Tumor site[a]	Locus	Nearby gene(s)	rsID lead variant	Chr.	Position (build 37)	Alleles (risk/ other)	RAF (%)	OR	95% CI	*P*	r^2	I^2	P_{het}	*N* cases	*N* controls
Colon	1p31.1	*PTGER3*	rs3124454	1	71,040,166	G/T	58.1	1.07	1.04-1.09	1.4E-08	0.926	6.1	0.38	32,002	64,159
Left-sided	2q21.3	*LCT*	rs1446585	2	136,407,479	G/A	39.9	1.07	1.04-1.10	3.3E-08	1.121	43.7	0.11	30,588	64,159
Proximal colon	3p22.2	*MLH1*	rs1800734[b]	3	37,034,946	A/G	24.7	1.15	1.11-1.19	3.8E-18	1.008	43.8	0.11	15,706	64,159
Colon	3p21.2	*STAB1; TLR9; NISCH*	rs353548	3	52,269,491	G/A	95.3	1.15	1.10-1.21	1.3E-08	0.975	0	0.48	32,002	64,159
Left-sided	5q32	*CDX1*	rs2302274[b]	5	149,546,426	G/A	47.8	1.07	1.04-1.09	4.9E-09	1.008	3.8	0.39	30,588	64,159
Left-sided	7q32.3	*KLF14; LINC00513*	rs73161913[b]	7	130,607,779	G/A	94.3	1.16	1.10-1.22	1.3E-09	0.975	0	0.79	30,588	64,159
Left-sided	10q23.31	*PANK1; KIF20B*	rs7071258[b]	10	91,574,624	A/G	21.6	1.08	1.05-1.11	8.4E-09	0.993	0	0.71	30,588	64,159
Rectal	14q22.1	*PYGL; NIN; ABHD12B*	rs28611105[b]	14	51,359,658	G/T	21.5	1.11	1.07-1.15	4.7E-09	0.983	50.5	0.07	16,212	64,159
Proximal colon	14q32.12	*RIN3*	rs61975764	14	93,014,929	G/A	55.3	1.08	1.05-1.11	2.8E-08	0.987	0	0.71	15,706	64,159
Proximal colon	14q32.2	*BCL11B*	rs80158569[b]	14	99,782,937	A/G	7.5	1.18	1.12-1.24	8.6E-11	0.899	29.9	0.21	15,706	64,159
Left-sided	19p13.3	*STK11; SBNO2*	rs62131228	19	1,157,642	G/A	98.1	1.28	1.17-1.40	2.4E-08	0.788	0	0.95	29,632	63,385
Left-sided	20q13.31	*BMP7*	rs6014965[b]	20	55,831,203	A/G	55.4	1.07	1.04-1.09	4.5E-09	0.995	10.5	0.35	30,588	64,159
Colon	22q13.31	*FAM118A; FBLN1*	rs736037	22	45,724,999	A/G	28.6	1.07	1.04-1.09	2.8E-08	1.015	0	0.74	32,002	64,159

Lead variant is the most significant variant at the locus. Reference SNP cluster ID (rsID) based on NCBI dbSNP Build 152. Alleles are on the + strand. Chr., chromosome; RAF, risk allele frequency; OR, odds (log-additive) ratio estimate for the risk allele; 95% CI, 95% confidence interval. 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 $N_{eff}=4/(1/N_{cases} + 1/N_{controls})$. The I^2 statistic measures heterogeneity on a scale of 0-100%. P_{het} is the *P*-value from Cochran's Q test for heterogeneity. [a]Colon: proximal colon + distal colon + colon, unspecified site; Left-sided: distal colon + rectal. Details of tumor site definitions including ICD-9 codes are given in the Methods section and Supplementary Materials. [b]Variant attained Bonferroni-adjusted genome-wide significance (5E-08/5 = 1E-08), corrected for the number of CRC case subgroups analyzed.

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

To gain insight into the molecular mechanisms underlying the new association signals, and to identify candidate causal variants and the most likely target gene(s), we annotated signals with functional and regulatory genomic annotations, assessed colocalization with expression quantitative trait loci (eQTLs), and performed literature-based gene prioritization. Notable and strong candidate causal variants and target genes are summarized here. Full results for all new signals are given in supplementary tables 4 and 5.

At the *MLH1* gene promoter region on 3p22.2, associated to proximal colon cancer risk, previous studies have reported strong and robust associations between the common SNP rs1800734, and sporadic CRC cases with high microsatellite instability (MSI-H) status.[19,20] 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 tumors.[20] The mechanisms of *MLH1* promoter hypermethylation and subsequent gene silencing may account for most sporadic CRC tumors with defective DNA MMR and MSI-H.[21]

At the highly localized, strongly proximal colon-specific association signal on 14q32.2, the lead SNP rs80158569 is located in a colonic crypt enhancer region and overlaps with multiple transcription factor binding sites, making it a strong candidate causal variant. The nearby gene *BCL11B* encodes a transcription factor that is required for normal T cell development,[22,23] and that has been identified as a SWI/SNF complex subunit.[24] *BCL11B* acts as a haploinsufficient tumor suppressor in T-cell acute lymphoblastic leukemia (T-ALL).[25,26] Experimental work reported by Sakamaki *et al.* suggests that impairment of Bcl11b promotes intestinal tumorigenesis in mice and humans through deregulation of the β-catenin pathway.[27]

At locus 14q32.12, lead SNP rs61975764 showed the strongest evidence of statistical association in the proximal colon analysis and attenuated effects for the other CRC tumor locations. Genotype-Tissue Expression (GTEx) data show that rs61975764 is an eQTL for gene Ras And Rab Interactor 3 (*RIN3*) in transverse colon tissue, the risk allele G being associated with decreased expression. RIN3 functions as a RAB5 and RAB31 guanine nucleotide exchange factor involved in endocytosis.[28,29]

At locus 5q32, one of six loci identified in the left-sided CRC analysis, 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. 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.[30] Consistent with this view, CDX1 has been shown to inhibit human colon cancer cell proliferation by blocking β-catenin/T-cell factor transcriptional activity.[31]

In a region of extensive LD on locus 2q21.1, lead SNP rs1446585, associated with left-sided CRC, is in strong LD with the functional SNP rs4988235 (LD $r^2 = 0.854$) in the *cis*-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 *P*-value for functional SNP rs4988235 when assuming an additive model was 7.0×10^{-7}. 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.[32] The protective effect conferred by the lactase persistence genotype is likely mediated by dairy products and calcium which are known protective factors for CRC.[33] Of note, the CC genotype has also been associated with a lower body mass index (BMI),[34] presumably because of the nutritional advantage associated with lactase persistence. Since this is a dominant trait with the rs4988235

CC genotype defining lactose intolerance, we also tested left-sided CRC association for these variants assuming a dominant model. Consistent with 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 the functional SNP rs4988235, the OR estimate for having genotype CC versus CT or TT, and left-sided CRC was 1.14 (95% CI: 1.09-1.19). Because this region has been under strong selective pressure, it is particularly prone to population stratification and follow-up studies are therefore warranted.[35] However, the fact that we included genotype principal components in the models for all analyzed sample sets, and that the association shows a consistent direction of effect across sample sets (supplementary table 6), suggest that this result is not driven by population stratification.

Candidate genes at the left-sided CRC risk loci 7q32.2 and 20q13.31 are involved in TGF-β signaling. At 7q32.3, the Krüppel-like factor 14 (*KLF14*) gene 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-beta receptor II (*TGFBR2*) gene.[36] A *cis*-eQTL for *KLF14*, that is uncorrelated with our lead SNP rs73161913, acts as a master regulator related to multiple metabolic phenotypes,[37,38] and an independent variant in this region has been associated to basal cell carcinoma.[39] For both reported associations, the effects depended on the parent-of-origin of the risk alleles. The association with metabolic phenotypes also depended on sex. We did not find any evidence for the presence of strong sex-dependent effects (males: OR=1.13, 95% CI=1.07-1.20, $P=4.4\times10^{-5}$; females: OR=1.17, 95% CI=1.09-1.25, $P=5.4\times10^{-6}$). Further investigation of this locus is warranted to analyze parent-of-origin effects on CRC risk, which is not possible in our dataset. At 20q13.31, 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.[40] BMP7

signaling in *TGFBR2*-deficient stromal cells promotes epithelial carcinogenesis through SMAD4-mediated signaling.[41] In CRC tumors, BMP7 expression correlates with parameters of pathological aggressiveness such as liver metastasis and poor prognosis.[42]

On 14q22.1, the single locus identified only in the rectal cancer analysis, GTEx data show that, in gastrointestinal tissues, colocalizes with a *cis*-eQTL co-regulating expression of genes *PYGL*, *ABHD12B*, and *NIN*. Glycogen Phosphorylase L (*PYGL*) is the strongest candidate. We recently identified and replicated an association between genetically predicted *PYGL* expression and CRC risk in a transcriptome-wide association study that used transverse colon tissue transcriptomes and genotypes from GTEx to construct prediction models.[43] Favaro *et al.* showed that this glycogen metabolism gene plays an important role in sustaining proliferation and preventing premature senescence in hypoxic cancer cells.[44] In different cancer cells lines, silencing of *PYGL*, expression of which is induced by exposure to hypoxia, led to increased glycogen accumulation and increased reactive oxygen species levels that contributed to p53-dependent induction of senescence and impaired tumorigenesis.[44]

At new locus 1p31.1, identified in the analysis for colon cancer, *PTGER3* 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-κβ 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.[45] Prostaglandin E2 (PGE2) is required for the activation of β-catenin by Wnt in stem cells,[46] and promotes colon cancer cell growth.[47] Prostaglandin E Receptor 3 plays an important role in suppression of cell growth and its downregulation was shown to enhance colon carcinogenesis.[48] Hypermethylation may contribute to its downregulation in colon cancer.[48]

Risk heterogeneity between tumor anatomical sublocations

Multinomial logistic regression modeling 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_{het} >0.05). For 51 of those variants, 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 (supplementary tables 2, 3 and 7; supplementary figure 5).

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

For 17 out of the 48 variants with P_{het} <0.05, the best-fitting model supported an effect limited to left-sided CRC (figure 2 and supplementary tables 3 and 7). Of these 17 variants, six were in the list of variants with the strongest evidence of heterogeneity (P_{het} <4.55×10^{-4}), including the following previously reported loci: *C11orf53-COLCA1-COLCA2* on 11q23.1 (P_{het}=6.0×10^{-14}), *APC* on 5q22.2 (P_{het}=2.3×10^{-10}), *GATA3* on 10p14 (P_{het}=1.7×10^{-8}), *CTNNB1* on 3p22.1 (P_{het}=9.8×10^{-8}), *RAB40B-METRLN* on 17q25.3 (P_{het}=3.6×10^{-6}), and *CDKN1A* on 6p21.2 (P_{het}=1.6×10^{-4}). Inspection of forest plots and association evidence also suggest stronger risk effects for left-sided tumors for the following additional five known loci: *TET2* on 4q24, *VTI1A* on 10q25.2, two independent signals near *POLD3* on 11q13.4, and *BMP4* on 14q22.2.

For 5 out of the 48 variants with P_{het} <0.05, a model with association with colon cancer risk, but no association with rectal cancer risk, provided the best fit (supplementary 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 suggest stronger risk effects for colon tumors for one of two independent signals near *PTPN1* on 20q13.13.

Evidence from the three approaches (figure 1; supplementary tables 3 and 7) indicates that only two loci are strongly proximal colon cancer-specific: *MLH1* on 3p22.2 ($P_{het}=5.4\times10^{-19}$), and *BCL11B* ($P_{het}=1.5\times10^{-5}$) on 14q32.2. Finally, for only 1 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.

DISCUSSION

It has long been recognized that sporadic CRCs arising in different anatomical segments of the colorectum differ in age- and sex-specific incidence rates, clinical, pathological and tumor molecular features. However, our understanding of the etiological factors underlying these medically important differences has remained scarce. This study aimed to examine whether the contribution of common germline genetic variants to sporadic CRC carcinogenesis differs by anatomical sublocation. The large sample size comprising 112,373 European-ancestry CRC cases and controls provided adequate statistical power to discover new loci and genetic variants with risk effects limited to tumors for certain anatomical subsites, and to compare allelic effect sizes of genetic variants across anatomical subsites.

Our CRC case subgroup meta-analyses identified 13 additional genome-wide significant CRC risk loci that, due to the presence of substantial allelic effect heterogeneity between anatomical subsites, were not detected in larger previously published GWAS for overall CRC risk.[8,9] In fact, the only way to discover certain loci and genetic risk variants with case subgroup-specific allelic effects is via analysis of homogeneous case subgroups.[49] For example, *P*-values for rs1800734 and rs80158569 were ~18 and ~5 powers of ten, respectively, more significant in the proximal colon analysis compared to in our overall

CRC analysis. While follow-up laboratory studies are needed to uncover the causal variant(s), the biological mechanism and the target gene, multiple lines of evidence support strong candidate target genes at many of the newly identified genome-wide significant loci, including genes *PTGER3*, *LCT*, *MLH1*, *CDX1*, *KLF14*, *PYGL*, *RIN3*, *BCL11B*, and *BMP7*.

Previous GWASs had already reported allelic effect heterogeneity between tumor sites, including for 10p14, 11q23, and 18q21 but only contrasted colon and rectal tumors, without distinguishing between proximal and distal colon.[50,51] The sample size and timing of the present study enabled a systematic characterization of heterogeneity of allelic effects between more primary tumor anatomical sublocations, and for a much expanded catalog of CRC risk variants. Our analysis revealed substantial and previously unappreciated allelic effect heterogeneity between proximal and distal CRC. The results further suggest that distal colon and rectal cancer have very similar germline genetic etiologies. Our findings at several loci are consistent with CRC tumor molecular studies. The consensus molecular subtypes (CMSs), which are based on tumor gene expression profiles, are differentially distributed between proximal and distal CRCs. The canonical CMS (CMS2) is highly enriched in distal CRC (56% versus 26% for proximal CRC) and is characterized by strong upregulation of Wnt downstream targets.[52] We found that risk 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 signaling pathway, were confined to distal CRC risk, suggesting that the antagonistic Wnt and Hedgehog signaling pathways may contribute more to the development of distal CRC tumors.

The precise intrinsic or extrinsic effect modifiers explaining the observed allelic effect heterogeneity between anatomical subsites remain unknown and further research is needed. Short-chain fatty acids (SCFAs), in particular butyrate, produced by microbiota through fermentation of dietary fiber in the human colon may be involved. Concentrations of butyrate, which plays a multifaceted antitumorigenic role in maintaining gut homeostasis, are much higher in the proximal colon.[53] Moreover, the known

chemopreventive role of butyrate may involve the modulation of signaling pathways including TGF-β and Wnt.[54] This may contribute to possible differences between anatomical subsites in colorectal crypt cellular dynamics with increased stem cell cycling in the distal colorectum promoting growth of precancerous conventional adenomas.

One limitation of our study is that we have not performed discovery GWAS analyses of CRC case subgroups based on more detailed anatomical locations beyond proximal colon, distal colon and rectum. However, given our current total sample size, such analyses would inevitably result in reduced statistical power for new discovery owing to the reduced sample sizes and the aggravated multiple testing burden. As another limitation, our study was based on subjects of European-ancestry and it remains to be determined whether findings are generalizable to other ancestry groups.

In conclusion, germline genetic data support the idea that proximal and distal colorectal cancer have partly distinct etiologies. Our results also demonstrate that future CRC germline genetic studies should take into consideration the differences between primary tumor anatomical subsites. A better understanding of the differing carcinogenic mechanisms and neoplastic transformation risk in the proximal and distal colorectum can inform the development of novel precision treatment and precision prevention strategies through the discovery of novel drug targets and repurposable drug candidates for chemoprevention and cancer treatment, and improved individualized screening recommendations based on risk prediction models that incorporate tumor anatomical subsite.

FUNDING AND ACKNOWLEDGEMENTS

Funding statements and acknowledgements are given in the supplemental text.

FIGURE LEGENDS

Figure 1. Primary tumor site-specific associations for the lead SNPs of the 13 CRC risk loci not reported in previous GWAS. The forest plots show the (log-additive) odds ratio estimates together with 95% confidence intervals. For clarity, this figure only shows results for the proximal colon, distal colon, and rectal cancer case subgroup analyses.

Figure 2. Loci showing association with risk of distal colorectal cancer (i.e., distal colon + rectal), but attenuated or no evidence for association with proximal colon cancer risk. The forest plots show the (log-additive) odds ratio estimates together with 95% confidence intervals from the GWAS meta-analyses of case subgroups defined by primary tumor 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 supplementary table 2. P_{het} is the *P*-value from a test for heterogeneity of effects across tumor subsites.

REFERENCES

1 American Cancer Society. Cancer Statistics Center. http://cancerstatisticscenter.cancer.org (accessed 20 April 2020).
2 Bufill JA. Colorectal cancer: evidence for distinct genetic categories based on proximal or distal tumor location. *Ann Intern Med* 1990;**113**:779–88. doi:10.7326/0003-4819-113-10-779
3 Iacopetta B. Are there two sides to colorectal cancer? *Int J Cancer* 2002;**101**:403–8. doi:10.1002/ijc.10635
4 Carethers JM. One colon lumen but two organs. *Gastroenterology* 2011;**141**:411–2. doi:10.1053/j.gastro.2011.06.029
5 Yamauchi M, Lochhead P, Morikawa T, *et al.* Colorectal cancer: a tale of two sides or a continuum? *Gut* 2012;**61**:794–7. doi:10.1136/gutjnl-2012-302014
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. doi:10.1056/NEJM200007133430201
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. doi:10.1093/jnci/djy099
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. doi:10.1038/s41588-018-0286-6
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. doi:10.1038/s41467-019-09775-w
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. doi:10.1053/j.gastro.2018.11.066
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.e24. doi:10.1053/j.gastro.2012.12.020
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. doi:10.1038/ncomms8138
13 McCarthy S, Das S, Kretzschmar W, *et al.* A reference panel of 64,976 haplotypes for genotype imputation. *Nat Genet* 2016;**48**:1279–83. doi:10.1038/ng.3643
14 Delaneau O, Howie B, Cox AJ, *et al.* Haplotype estimation using sequencing reads. *Am J Hum Genet* 2013;**93**:687–96. doi:10.1016/j.ajhg.2013.09.002
15 Das S, Forer L, Schönherr S, *et al.* Next-generation genotype imputation service and methods. *Nat Genet* 2016;**48**:1284–7. doi:10.1038/ng.3656
16 Willer CJ, Li Y, Abecasis GR. METAL: fast and efficient meta-analysis of genomewide association scans. *Bioinformatics* 2010;**26**:2190–1. doi:10.1093/bioinformatics/btq340
17 Burnham KP, Anderson DR. Multimodel Inference: Understanding AIC and BIC in Model Selection. *Sociol Methods Res* 2004;**33**:261–304. doi:10.1177/0049124104268644
18 Yee TW. The VGAM package for categorical data analysis. *J Stat Softw* 2010;**32**. doi:10.18637/jss.v032.i10
19 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

20 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
21 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.
22 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
23 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
24 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
25 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
26 Neumann M, Vosberg S, Schlee C, *et al.* Mutational spectrum of adult T-ALL. *Oncotarget* 2015;**6**:2754–66. doi:10.18632/oncotarget.2218
27 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
28 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
29 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
30 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
31 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
32 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
33 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*
34 Kettunen J, Silander K, Saarela O, *et al.* European lactase persistence genotype shows evidence of association with increase in body mass index. *Hum Mol Genet* 2010;**19**:1129–36. doi:10.1093/hmg/ddp561
35 Campbell CD, Ogburn EL, Lunetta KL, *et al.* Demonstrating stratification in a European American population. *Nat Genet* 2005;**37**:868–72. doi:10.1038/ng1607
36 Truty MJ, Lomberk 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

37 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

38 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

39 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

40 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

41 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. doi:10.1158/1541-7786.MCR-18-0120

42 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

43 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

44 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

45 Jänne PA, Mayer RJ. Chemoprevention of colorectal cancer. *N Engl J Med* 2000;**342**:1960–8. doi:10.1056/NEJM200006293422606

46 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

47 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

48 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

49 Traylor M, Markus H, Lewis CM. Homogeneous case subgroups increase power in genetic association studies. *Eur J Hum Genet* 2015;**23**:863–9. doi:10.1038/ejhg.2014.194

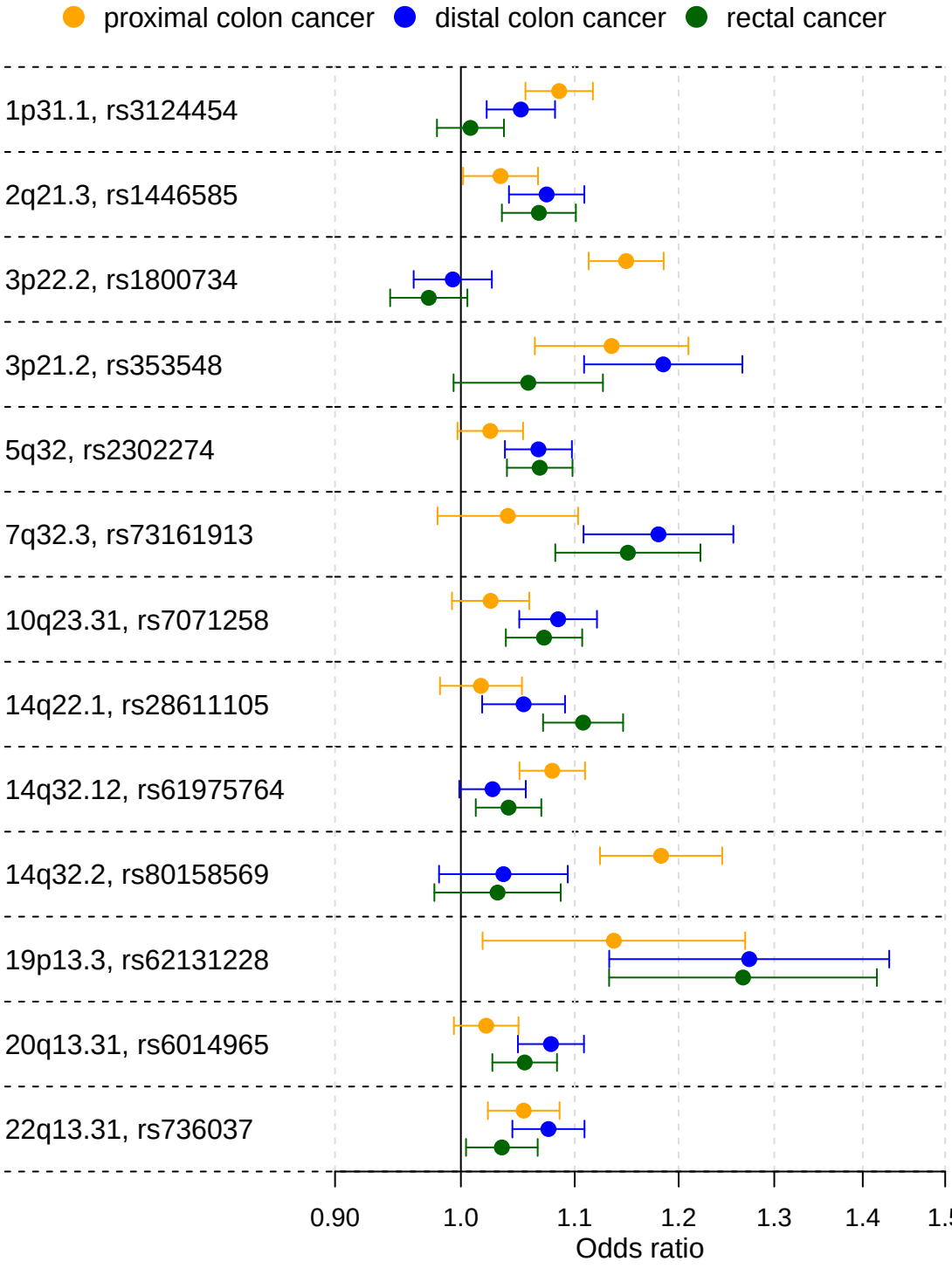
50 Tenesa A, Farrington SM, Prendergast JGD, *et al.* Genome-wide association scan identifies a colorectal cancer susceptibility locus on 11q23 and replicates risk loci at 8q24 and 18q21. *Nat Genet* 2008;**40**:631–7. doi:10.1038/ng.133

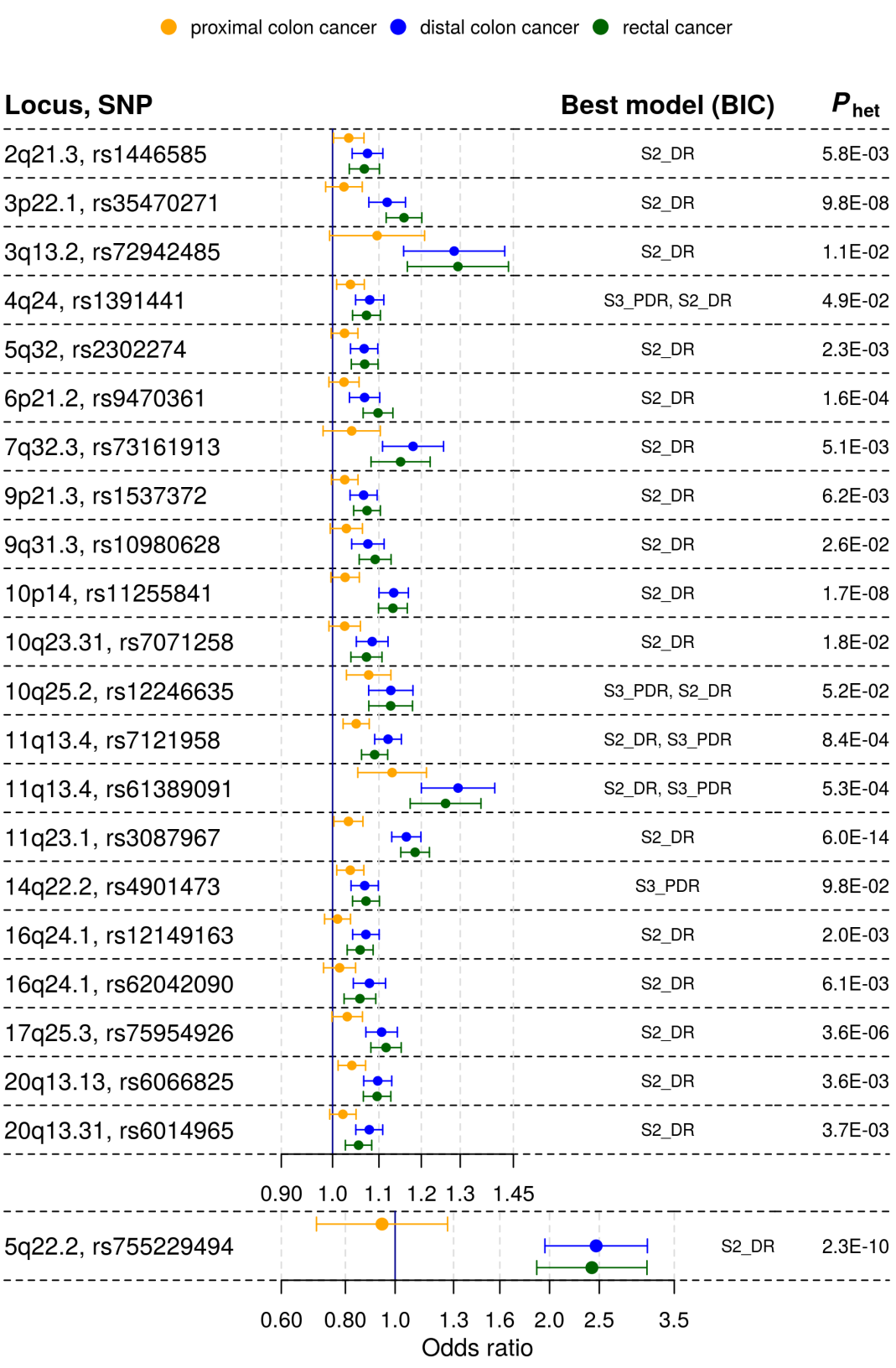
51 Tomlinson IPM, Webb E, Carvajal-Carmona L, *et al.* A genome-wide association study identifies colorectal cancer susceptibility loci on chromosomes 10p14 and 8q23.3. *Nat Genet* 2008;**40**:623–30. doi:10.1038/ng.111

52 Guinney J, Dienstmann R, Wang X, *et al.* The consensus molecular subtypes of colorectal cancer. *Nat Med* 2015;**21**:1350–6. doi:10.1038/nm.3967

53 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. doi:10.1016/B978-0-12-800100-4.00003-9

54 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**. doi:10.3390/nu9121348


proximal colon cancer
distal colon cancer
rectal cancer
1p31.1, rs3124454
2q21.3, rs1446585
3p22.2, rs1800734
3p21.2, rs353548
5q32, rs2302274
7q32.3, rs73161913
10q23.31, rs7071258
14q22.1, rs28611105
14q32.12, rs61975764
14q32.2, rs80158569
19p13.3, rs62131228
20q13.31, rs6014965
22q13.31, rs736037
0.90
1.0
1.1
1.2
1.3
1.4
1.5
Odds ratio


proximal colon cancer
distal colon cancer
rectal cancer
Locus, SNP
Best model (BIC)
P_{het}
2q21.3, rs1446585 S2_DR 5.8E-03
3p22.1, rs35470271 S2_DR 9.8E-08
3q13.2, rs72942485 S2_DR 1.1E-02
4q24, rs1391441 S3_PDR, S2_DR 4.9E-02
5q32, rs2302274 S2_DR 2.3E-03
6p21.2, rs9470361 S2_DR 1.6E-04
7q32.3, rs73161913 S2_DR 5.1E-03
9p21.3, rs1537372 S2_DR 6.2E-03
9q31.3, rs10980628 S2_DR 2.6E-02
10p14, rs11255841 S2_DR 1.7E-08
10q23.31, rs7071258 S2_DR 1.8E-02
10q25.2, rs12246635 S3_PDR, S2_DR 5.2E-02
11q13.4, rs7121958 S2_DR, S3_PDR 8.4E-04
11q13.4, rs61389091 S2_DR, S3_PDR 5.3E-04
11q23.1, rs3087967 S2_DR 6.0E-14
14q22.2, rs4901473 S3_PDR 9.8E-02
16q24.1, rs12149163 S2_DR 2.0E-03
16q24.1, rs62042090 S2_DR 6.1E-03
17q25.3, rs75954926 S2_DR 3.6E-06
20q13.13, rs6066825 S2_DR 3.6E-03
20q13.31, rs6014965 S2_DR 3.7E-03
0.90 1.0 1.1 1.2 1.3 1.45
5q22.2, rs755229494 S2_DR 2.3E-10
0.60 0.80 1.0 1.3 1.6 2.0 2.5 3.5
Odds ratio