Discovery of common and rare genetic risk variants for colorectal cancer

Jeroen R Huyghe[1]*, Stephanie A Bien[1]*, Tabitha A Harrison[1]*, Hyun Min Kang[2], Sai Chen[2], Stephanie L Schmit[3], David V Conti[4], Conghui Qu[1], Jihyoun Jeon[5], Christopher K Edlund[4], Peyton Greenside[6], Michael Wainberg[7], Fredrick R Schumacher[8], Joshua D Smith[9], David M Levine[10], Sarah C Nelson[10], Nasa A Sinnott-Armstrong[11], Demetrius Albanes[12], M Henar Alonso[13–15], Kristin Anderson[16], Coral Arnau-Collell[17], Volker Arndt[18], Christina Bamia[19,20], Barbara L Banbury[1], John A Baron[21], Sonja I Berndt[12], Stéphane Bézieau[22], D Timothy Bishop[23], Juergen Boehm[24], Heiner Boeing[25], Hermann Brenner[18,26,27], Stefanie Brezina[28], Stephan Buch[29], Daniel D Buchanan[30–32], Andrea Burnett-Hartman[33], Katja Butterbach[18], Bette J Caan[34], Peter T Campbell[35], Christopher S Carlson[1,36], Sergi Castellví-Bel[17], Andrew T Chan[37–42], Jenny Chang-Claude[43,44], Stephen J Chanock[12], Maria-Dolores Chirlaque[14,45], Sang Hee Cho[46], Charles M Connolly[1], Amanda J Cross[47,48] Katarina Cuk[18], Keith R Curtis[1], Albert de la Chapelle[49], Kimberly F Doheny[50], David Duggan[51], Douglas F Easton[52,53], Sjoerd G Elias[54], Faye Elliott[23], Dallas R English[55,56], Edith JM Feskens[57], Jane C Figueiredo[58,59], Rocky Fischer[60], Liesel M FitzGerald[56,61], David Forman[62], Manish Gala[37,39], Steven Gallinger[63], W James Gauderman[4], Graham G Giles[55,56], Elizabeth Gillanders[64], Jian Gong[1], Phyllis J Goodman[65], William M Grady[66], John S Grove[67], Andrea Gsur[28], Marc J Gunter[68], Robert W Haile[69], Jochen Hampe[29], Heather Hampel[70], Sophia Harlid[71], Richard B Hayes[72], Philipp Hofer[28], Michael Hoffmeister[18], John L Hopper[55,73], Wan-Ling Hsu[10], Wen-Yi Huang[12], Thomas J Hudson[74], David J Hunter[41,75], Gemma Ibañez-Sanz[13,76,77], Gregory E Idos[4], Roxann Ingersoll[50], Rebecca D Jackson[78], Eric J Jacobs[35], Mark A Jenkins[55], Amit D Joshi[39,41], Corinne E Joshu[79], Temitope O Keku[80], Timothy J Key[81], Hyeong Rok Kim[82], Emiko Kobayashi[1], Laurence N Kolonel[83], Charles Kooperberg[1], Tilman Kühn[43], Sébastien Küry[22], Sun-Seog Kweon[84,85], Susanna C Larsson[86], Cecelia A Laurie[10], Loic Le Marchand[67], Suzanne M Leal[87], Soo Chin Lee[88,89], Flavio Lejbkowicz[90–92], Mathieu Lemire[74], Christopher I Li[1], Li Li[93], Wolfgang Lieb[94], Yi Lin[1], Annika Lindblom[95,96], Noralane M Lindor[97], Hua Ling[50], Tin L Louie[10], Satu Männistö[98], Sanford D Markowitz[99], Vicente Martín[14,100], Giovanna Masala[101], Caroline E McNeil[102], Marilena Melas[4], Roger L Milne[55,56], Lorena Moreno[17], Neil Murphy[68], Robin Myte[71], Alessio Naccarati[103,104], Polly A Newcomb[1,36], Kenneth Offit[105,106], Shuji Ogino[40,41,107,108], N Charlotte Onland-Moret[54], Barbara Pardini[104,109], Patrick S Parfrey[110], Rachel Pearlman[70], Vittorio Perduca[111,112], Paul D P Pharoah[52], Mila Pinchev[91], Elizabeth A Platz[79], Ross L Prentice[1], Elizabeth Pugh[50], Leon Raskin[113], Gad Rennert[91,92,114], Hedy S Rennert[91,92,114], Elio Riboli[115], Miguel Rodríguez-Barranco[14,116], Jane Romm[50], Lori C Sakoda[1,117], Clemens Schafmayer[118], Robert E Schoen[119], Daniela Seminara[64], Mitul Shah[53], Tameka Shelford[50], Min-Ho Shin[84], Katerina Shulman[120], Sabina Sieri[121], Martha L Slattery[122], Melissa C Southey[123], Zsofia K Stadler[124], Christa Stegmaier[125], Yu-Ru Su[1], Catherine M Tangen[65], Stephen N Thibodeau[126], Duncan C Thomas[4], Sushma S Thomas[1], Amanda E Toland[127], Antonia Trichopoulou[19,20], Cornelia M Ulrich[24], David J Van Den Berg[4], Franzel JB van Duijnhoven[57], Bethany Van Guelpen[71], Henk van Kranen[128], Joseph Vijai[124], Kala Visvanathan[79], Pavel Vodicka[103,129,130], Ludmila Vodickova[103,129,130], Veronika Vymetalkova[103,129,130], Korbinian Weigl[18,27,131], Stephanie J Weinstein[12], Emily White[1], Aung Ko Win[32,55], C Roland Wolf[132], Alicja Wolk[86,133], Michael O Woods[134], Anna H Wu[4], Syed H Zaidi[74], Brent W Zanke[135], Qing Zhang[136], Wei Zheng[137], Peter C Scacheri[138], John D Potter[1], Michael C Bassik[11], Anshul Kundaje[7,11], Graham Casey[139], Victor Moreno[13–15,77], Goncalo R Abecasis[2], Deborah A Nickerson[9]§, Stephen B Gruber[4]§, Li Hsu[1,10]§, Ulrike Peters[1,36]§

1. Public Health Sciences Division, Fred Hutchinson Cancer Research Center, Seattle, Washington, USA.
2. Department of Biostatistics and Center for Statistical Genetics, University of Michigan, Ann Arbor, Michigan, USA.
3. Department of Cancer Epidemiology, H. Lee Moffitt Cancer Center and Research Institute, Tampa, Florida, USA.
4. Department of Preventive Medicine, USC Norris Comprehensive Cancer Center, Keck School of Medicine, University of Southern California, Los Angeles, California, USA.
5. Department of Epidemiology, University of Michigan, Ann Arbor, Michigan, USA.
6. Biomedical Informatics Program, Stanford University, Stanford, California, USA.
7. Department of Computer Science, Stanford University, Stanford, California, USA.
8. Department of Population and Quantitative Health Sciences, Case Western Reserve University, Cleveland, Ohio, USA.
9. Department of Genome Sciences, University of Washington, Seattle, Washington, USA.
10. Department of Biostatistics, University of Washington, Seattle, Washington, USA.
11. Department of Genetics, Stanford University, Stanford, California, USA.
12. Division of Cancer Epidemiology and Genetics, National Cancer Institute, National Institutes of Health, Bethesda, Maryland, USA.
13. Cancer Prevention and Control Program, Catalan Institute of Oncology-IDIBELL, L'Hospitalet de Llobregat, Barcelona, Spain.
14. CIBER de Epidemiología y Salud Pública (CIBERESP), Madrid, Spain.
15. Department of Clinical Sciences, Faculty of Medicine, University of Barcelona, Barcelona, Spain.
16. Division of Epidemiology and Community Health, University of Minnesota, Minneapolis, Minnesota, USA.
17. 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.
18. Division of Clinical Epidemiology and Aging Research, German Cancer Research Center (DKFZ), Heidelberg, Germany.
19. Hellenic Health Foundation, Athens, Greece.
20. 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.
21. Department of Medicine, University of North Carolina School of Medicine, Chapel Hill, North Carolina, USA.
22. Service de Génétique Médicale, Centre Hospitalier Universitaire (CHU) Nantes, Nantes, France.
23. Leeds Institute of Medical Research at St James's, University of Leeds, Leeds, UK.
24. Huntsman Cancer Institute and Department of Population Health Sciences, University of Utah, Salt Lake City, Utah, USA.
25. Department of Epidemiology, German Institute of Human Nutrition (DIfE), Potsdam-Rehbrücke, Germany.

26. Division of Preventive Oncology, German Cancer Research Center (DKFZ) and National Center for Tumor Diseases (NCT), Heidelberg, Germany.
27. German Cancer Consortium (DKTK), German Cancer Research Center (DKFZ), Heidelberg, Germany.
28. Institute of Cancer Research, Department of Medicine I, Medical University of Vienna, Vienna, Austria.
29. Department of Medicine I, University Hospital Dresden, Technische Universität Dresden (TU Dresden), Dresden, Germany.
30. Colorectal Oncogenomics Group, Department of Clinical Pathology, The University of Melbourne, Parkville, Victoria, Australia.
31. University of Melbourne Centre for Cancer Research, Victorian Comprehensive Cancer Centre, Parkville, Victoria, Australia.
32. Genomic Medicine and Family Cancer Clinic, Royal Melbourne Hospital, Parkville, Victoria, Australia.
33. Institute for Health Research, Kaiser Permanente Colorado, Denver, Colorado, USA.
34. Division of Research, Kaiser Permanente Medical Care Program, Oakland, California, USA.
35. Behavioral and Epidemiology Research Group, American Cancer Society, Atlanta, Georgia, USA.
36. Department of Epidemiology, University of Washington, Seattle, Washington, USA.
37. Division of Gastroenterology, Massachusetts General Hospital and Harvard Medical School, Boston, Massachusetts, USA.
38. Channing Division of Network Medicine, Brigham and Women's Hospital and Harvard Medical School, Boston, Massachusetts, USA.
39. Clinical and Translational Epidemiology Unit, Massachusetts General Hospital and Harvard Medical School, Boston, Massachusetts, USA.
40. Broad Institute of Harvard and MIT, Cambridge, Massachusetts, USA.
41. Department of Epidemiology, Harvard T.H. Chan School of Public Health, Harvard University, Boston, Massachusetts, USA.
42. Department of Immunology and Infectious Diseases, Harvard T.H. Chan School of Public Health, Harvard University, Boston, Massachusetts, USA.
43. Division of Cancer Epidemiology, German Cancer Research Center (DKFZ), Heidelberg, Germany.
44. Cancer Epidemiology Group, University Medical Centre Hamburg-Eppendorf, University Cancer Centre Hamburg (UCCH), Hamburg, Germany.
45. Department of Epidemiology, Regional Health Council, IMIB-Arrixaca, Murcia University, Murcia, Spain.
46. Department of Hematology-Oncology, Chonnam National University Hospital, Hwasun, South Korea.
47. Department of Epidemiology and Biostatistics, Imperial College London, London, UK.
48. Department of Surgery and Cancer, Imperial College London, London, UK. .
49. Department of Cancer Biology and Genetics and the Comprehensive Cancer Center, The Ohio State University, Columbus, Ohio, USA.
50. Center for Inherited Disease Research (CIDR), Institute of Genetic Medicine, Johns Hopkins University, Baltimore, Maryland, USA.

51. Translational Genomics Research Institute - An Affiliate of City of Hope, Phoenix, Arizona, USA.
52. Department of Public Health and Primary Care, University of Cambridge, Cambridge, UK.
53. Centre for Cancer Genetic Epidemiology, Department of Oncology, University of Cambridge, Cambridge, UK.
54. Julius Center for Health Sciences and Primary Care, University Medical Center Utrecht, Utrecht University, Utrecht, The Netherlands.
55. Centre for Epidemiology and Biostatistics, Melbourne School of Population and Global Health, The University of Melbourne, Melbourne, Victoria, Australia.
56. Cancer Epidemiology and Intelligence Division, Cancer Council Victoria, Melbourne, Victoria, Australia.
57. Division of Human Nutrition and Health, Wageningen University and Research, Wageningen, The Netherlands.
58. Department of Medicine, Samuel Oschin Comprehensive Cancer Institute, Cedars-Sinai Medical Center, Los Angeles, California, USA.
59. Department of Preventive Medicine, Keck School of Medicine, University of Southern California, Los Angeles, California, USA.
60. University of Michigan Comprehensive Cancer Center, Ann Arbor, Michigan, USA.
61. Menzies Institute for Medical Research, University of Tasmania, Hobart, Tasmania, Australia.
62. International Agency for Research on Cancer, World Health Organization, Lyon, France.
63. Lunenfeld Tanenbaum Research Institute, Mount Sinai Hospital, University of Toronto, Toronto, Ontario, Canada.
64. Division of Cancer Control and Population Sciences, National Cancer Institute, Bethesda, Maryland, USA.
65. SWOG Statistical Center, Fred Hutchinson Cancer Research Center, Seattle, Washington, USA.
66. Clinical Research Division, Fred Hutchinson Cancer Research Center, Seattle, Washington, USA.
67. University of Hawaii Cancer Research Center, Honolulu, Hawaii, USA.
68. Nutrition and Metabolism Section, International Agency for Research on Cancer, World Health Organization, Lyon, France.
69. Division of Oncology, Department of Medicine, Stanford University, Stanford, California, USA.
70. Division of Human Genetics, Department of Internal Medicine, The Ohio State University Comprehensive Cancer Center, Columbus, Ohio, USA.
71. Department of Radiation Sciences, Oncology Unit, Umeå University, Umeå, Sweden.
72. Division of Epidemiology, Department of Population Health, New York University School of Medicine, New York, New York, USA.
73. Department of Epidemiology, School of Public Health and Institute of Health and Environment, Seoul National University, Seoul, South Korea.
74. Ontario Institute for Cancer Research, Toronto, Ontario, Canada.
75. Nuffield Department of Population Health, University of Oxford, Oxford, UK.
76. Gastroenterology Department, Bellvitge University Hospital, L'Hospitalet de Llobregat, Barcelona, Spain.

77. Colorectal Cancer Group, ONCOBELL Program, Bellvitge Biomedical Research Institute-IDIBELL, Hospitalet de Llobregat, Barcelona, Spain.
78. Department of Medicine, Division of Endocrinology, Diabetes and Metabolism, The Ohio State University, Columbus, Ohio, USA.
79. Department of Epidemiology, Johns Hopkins Bloomberg School of Public Health, Johns Hopkins University, Baltimore, Maryland, USA.
80. Center for Gastrointestinal Biology and Disease, University of North Carolina, Chapel Hill, North Carolina, USA.
81. Cancer Epidemiology Unit, Nuffield Department of Population Health, University of Oxford, Oxford, UK.
82. Department of Surgery, Chonnam National University Hwasun Hospital and Medical School, Hwasun, Korea.
83. Office of Public Health Studies, University of Hawaii Manoa, Honolulu, Hawaii, USA.
84. Department of Preventive Medicine, Chonnam National University Medical School, Gwangju, Korea.
85. Jeonnam Regional Cancer Center, Chonnam National University Hwasun Hospital, Hwasun, Korea.
86. Institute of Environmental Medicine, Karolinska Institutet, Stockholm, Sweden.
87. Center for Statistical Genetics, Department of Molecular and Human Genetics, Baylor College of Medicine, Houston, Texas, USA.
88. Department of Haematology-Oncology, National University Cancer Institute, Singapore.
89. Cancer Science Institute of Singapore, National University of Singapore, Singapore.
90. The Clalit Health Services, Personalized Genomic Service, Carmel, Haifa, Israel.
91. Department of Community Medicine and Epidemiology, Lady Davis Carmel Medical Center, Haifa, Israel.
92. Clalit National Cancer Control Center, Haifa, Israel.
93. Center for Community Health Integration and Case Comprehensive Cancer Center, Case Western Reserve University, Cleveland, Ohio, USA.
94. Institute of Epidemiology, PopGen Biobank, Christian-Albrechts-University Kiel, Kiel, Germany.
95. Department of Clinical Genetics, Karolinska University Hospital, Stockholm, Sweden.
96. Department of Molecular Medicine and Surgery, Karolinska Institutet, Stockholm, Sweden.
97. Department of Health Science Research, Mayo Clinic, Scottsdale, Arizona, USA.
98. Department of Public Health Solutions, National Institute for Health and Welfare, Helsinki, Finland.
99. Departments of Medicine and Genetics, Case Comprehensive Cancer Center, Case Western Reserve University, and University Hospitals of Cleveland, Cleveland, Ohio, USA.
100. Biomedicine Institute (IBIOMED), University of León, León, Spain.
101. Cancer Risk Factors and Life-Style Epidemiology Unit, Institute of Cancer Research, Prevention and Clinical Network - ISPRO, Florence, Italy.
102. USC Norris Comprehensive Cancer Center, University of Southern California, Los Angeles, California, USA.
103. Department of Molecular Biology of Cancer, Institute of Experimental Medicine of the Czech Academy of Sciences, Prague, Czech Republic.
104. Italian Institute for Genomic Medicine (IIGM), Turin, Italy.

105. Clinical Genetics Service, Department of Medicine, Memorial Sloan Kettering Cancer Center, New York, New York, USA.
106. Department of Medicine, Weill Cornell Medical College, New York, New York, USA.
107. Program in MPE Molecular Pathological Epidemiology, Department of Pathology, Brigham and Women's Hospital, Harvard Medical School, Boston, Massachusetts, USA.
108. Department of Oncologic Pathology, Dana-Farber Cancer Institute, Boston, Massachusetts, USA.
109. Department of Medical Sciences, University of Turin, Turin, Italy.
110. The Clinical Epidemiology Unit, Memorial University Medical School, Newfoundland, Canada.
111. Laboratoire de Mathématiques Appliquées MAP5 (UMR CNRS 8145), Université Paris Descartes, Paris, France.
112. CESP (Inserm U1018), Facultés de Medicine Université Paris-Sud, UVSQ, Université Paris-Saclay, Gustave Roussy, Villejuif, France.
113. Division of Epidemiology, Vanderbilt Epidemiology Center, Vanderbilt University School of Medicine, Nashville, Tennessee, USA.
114. Ruth and Bruce Rappaport Faculty of Medicine, Technion-Israel Institute of Technology, Haifa, Israel.
115. School of Public Health, Imperial College London, London, UK.
116. Escuela Andaluza de Salud Pública. Instituto de Investigación Biosanitaria ibs.GRANADA. Hospitales Universitarios de Granada/Universidad de Granada, Granada, Spain.
117. Division of Research, Kaiser Permanente Northern California, Oakland, California, USA.
118. Department of General and Thoracic Surgery, University Hospital Schleswig-Holstein, Campus Kiel, Kiel, Germany.
119. Department of Medicine and Epidemiology, University of Pittsburgh Medical Center, Pittsburgh, Pennsylvania, USA.
120. Oncology Unit, Hillel Yaffe Medical Center, Hadera, Israel.
121. Epidemiology and Prevention Unit, Fondazione IRCCS Istituto Nazionale dei Tumori, Milan, Italy.
122. Department of Internal Medicine, University of Utah, Salt Lake City, Utah, USA.
123. Genetic Epidemiology Laboratory, Department of Pathology, The University of Melbourne, Melbourne, Australia.
124. Department of Medicine, Memorial Sloan Kettering Cancer Center, New York, New York, USA.
125. Saarland Cancer Registry, Saarbrücken, Germany.
126. Division of Laboratory Genetics, Department of Laboratory Medicine and Pathology, Mayo Clinic, Rochester, Minnesota, USA.
127. Departments of Cancer Biology and Genetics and Internal Medicine, Comprehensive Cancer Center, The Ohio State University, Columbus, Ohio, USA.
128. National Institute for Public Health and the Environment (RIVM), Bilthoven, The Netherlands.
129. Institute of Biology and Medical Genetics, First Faculty of Medicine, Charles University, Prague, Czech Republic.
130. Faculty of Medicine and Biomedical Center in Pilsen, Charles University, Pilsen, Czech Republic.

131. Medical Faculty, University of Heidelberg, Germany.
132. School of Medicine, University of Dundee, Dundee, Scotland.
133. Department of Surgical Sciences, Uppsala University, Uppsala, Sweden.
134. Memorial University of Newfoundland, Discipline of Genetics, St. John's, Canada.
135. Division of Hematology, University of Toronto, Toronto, Ontario, Canada.
136. Genomics Shared Resource, Fred Hutchinson Cancer Research Center, Seattle, Washington, USA.
137. Division of Epidemiology, Department of Medicine, Vanderbilt-Ingram Cancer Center, Vanderbilt Epidemiology Center, Vanderbilt University School of Medicine, Nashville, Tennessee, USA.
138. Department of Genetics and Genome Sciences, Case Western Reserve University School of Medicine, Case Comprehensive Cancer Center, Cleveland, Ohio, USA.
139. Center for Public Health Genomics, University of Virginia, Charlottesville, Virginia, USA.

*These authors contributed equally to this work.

§These authors jointly supervised this work.

Correspondence should be addressed to U.P. (upeters@fredhutch.org).

To further dissect the genetic architecture of colorectal cancer (CRC), we performed whole-genome sequencing of 1,439 cases and 720 controls, imputed discovered sequence variants and Haplotype Reference Consortium panel variants into genome-wide association study data, and tested for association in 34,869 cases and 29,051 controls. Findings were followed up in an additional 23,262 cases and 38,296 controls. We discovered a strongly protective 0.3% frequency variant signal at *CHD1*. In a combined meta-analysis of 125,478 individuals, we identified 40 new independent signals at $P<5\times10^{-8}$, bringing the number of known independent signals for CRC to approximately 100. New signals implicate lower-frequency variants, Krüppel-like factors, Hedgehog signaling, Hippo-YAP signaling, long noncoding RNAs, somatic drivers, and support a role of immune function. Heritability analyses suggest that CRC risk is highly polygenic, and larger, more comprehensive studies enabling rare variant analysis will improve understanding of underlying biology, and impact personalized screening strategies and drug development.

Colorectal cancer (CRC) is the fourth leading cancer-related cause of death worldwide[1] and presents a major public health burden. Up to 35% of inter-individual variability in CRC risk has been attributed to genetic factors[2,3]. Family-based studies have identified rare high-penetrance mutations in at least a dozen genes but, collectively, these account for only a small fraction of

310 familial risk[4]. Over the past decade, genome-wide association studies (GWAS) for sporadic
311 CRC, which constitutes the majority of cases, have identified approximately 60 association
312 signals at over 50 loci[5–22]. Yet, most of the genetic factors contributing to CRC risk remain
313 undefined. This severely hampers our understanding of biological processes underlying CRC. It
314 also limits CRC precision prevention, including individualized preventive screening
315 recommendations and development of cancer prevention drugs. The contribution of rare
316 variation to sporadic CRC is particularly poorly understood.
317
318 To expand the catalog of CRC risk loci and improve our understanding of rare variants, genes,
319 and pathways influencing sporadic CRC risk, and risk prediction, we performed the largest and
320 most comprehensive whole-genome sequencing (WGS) study and GWAS meta-analysis for
321 CRC to date, combining data from three consortia: the Genetics and Epidemiology of Colorectal
322 Cancer Consortium (GECCO), the Colorectal Cancer Transdisciplinary Study (CORECT), and
323 the Colon Cancer Family Registry (CCFR). Our study almost doubles the number of individuals
324 analyzed, incorporating GWAS results from >125,000 individuals, and substantially expands and
325 strengthens our understanding of biological processes underlying CRC risk.
326
327 **RESULTS**
328 **Study Overview**
329 We performed WGS of 1,439 CRC cases and 720 controls of European ancestry at low coverage
330 (3.8-8.6×). We detected, called, and estimated haplotype phase for 31.8 million genetic variants,
331 including 1.7 million short insertion-deletion variants (indels) (Online Methods). These data
332 include many rare variants not studied by GWAS. Based on other large-scale WGS studies
333 employing a similar design, we expected to have near-complete ascertainment of single
334 nucleotide variants (SNVs) with minor allele count (MAC) greater than five (minor allele
335 frequency (MAF) >0.1%), and high accuracy at heterozygous genotypes[23,24]. We tested 14.4
336 million variants with MAC ≥5 for CRC association using logistic regression (Online Methods)
337 but did not find any significant associations. To increase power to detect associations with rare
338 and low-frequency variants of modest effect, we imputed variants from the sequencing
339 experiment into 34,869 cases and 29,051 controls of predominantly European (91.7%) and East
340 Asian ancestry (8.3%) from 30 existing GWAS studies (Online Methods and **Supplementary**

Table 1). By design, two thirds of sequenced individuals were CRC cases, thereby enriching the panel for rare or low-frequency alleles that increase CRC risk. We contributed our sequencing data to the Haplotype Reference Consortium (HRC)[25] and imputed the 30 existing GWAS studies to the HRC panel, which comprises haplotypes for 32,488 individuals. Results of these GWAS meta-analyses (referred to as Stage 1 meta-analysis; Online Methods) informed the design of a custom Illumina array comprising the OncoArray, a custom array to identify cancer risk loci[26], and 15,802 additional variants selected based on Stage 1 meta-analysis results. We genotyped 12,007 cases and 12,000 controls of European ancestry with this custom array, and combined them with an additional 11,255 cases and 26,296 controls with GWAS data, resulting in a Stage 2 meta-analysis of 23,262 CRC cases and 38,296 controls (Online Methods, **Supplementary Fig. 1**, and **Supplementary Table 1**). Next, we performed a combined (Stage 1 + Stage 2) meta-analysis of up to 58,131 cases and 67,347 controls. This meta-analysis was based on the HRC-panel-imputed data because, given its large size, this panel results in superior imputation quality and enables accurate imputation of variants with MAFs as low as 0.1%[25]. Here, we report new association signals discovered through our custom genotyping experiment and replicating in Stage 2 at the Bonferroni significance threshold of $P < 7.8\times10^{-6}$ (Online Methods), as well as distinct association signals passing the genome-wide significance (GWS) threshold of $P < 5\times10^{-8}$ in the combined meta-analysis of up to 125,478 individuals.

CRC risk loci

In the combined meta-analysis, we identified 30 new CRC risk loci reaching GWS and >500kb away from previously reported CRC risk variants (**Table 1**; **Supplementary Fig. 2 and 3**). Twenty-two of these were represented on our custom genotyping panel, either by the lead variant (15 loci) or by a variant in linkage disequilibrium (LD) (7 loci; r^2>0.7). Of these 22 variants, eight attained the Bonferroni significance threshold in the Stage 2 meta-analysis (**Table 1**).

Among these eight loci is the first rare variant signal identified for sporadic CRC, involving five 0.3% frequency variants at 5q21.1, near genes *CHD1* and *RGMB*. SNP rs145364999, intronic to *CHD1*, had high quality genotyping (**Supplementary Fig. 4**). The variant was well imputed in the remaining sample sets (imputation quality r^2 ranged from 0.66 to 0.87; **Supplementary Table 2**) and there was no evidence of heterogeneity of effects (heterogeneity P=0.63;

Supplementary Table 2). The rare allele confers a strong protective effect (allelic odds ratio (OR)=0.52 in Stage 2; 95% confidence interval (CI)=0.40-0.68). Chromatin remodeling factor CHD1 provides an especially plausible candidate and has been shown to be a synthetically-essential gene[27] that is occasionally deleted in some cancers, but always retained in PTEN-deficient cancers[28]. The resulting mutually exclusive deletion pattern of *CHD1* and *PTEN* has been observed in prostate, breast, and CRC TCGA data[28]. We hypothesize that the rare allele confers a protective effect through lowering *CHD1* expression, which is required for nuclear factor-κβ (NF-κβ) pathway activation and growth in cancer cells driven by loss of the tumor suppressor *PTEN*[28]. However, we cannot rule out involvement of nearby candidate gene *RGMB* that encodes a co-receptor for bone morphogenetic proteins BMP2 and BMP4, both of which are linked to CRC risk through GWAS[9,11]. Additionally, RGMB has been shown to bind to PD-L2[29], a known ligand of PD-1, an immune checkpoint blockade inhibitor targeted by cancer immunotherapy[30].

The vast majority of new association signals involve common variants. We found associations near strong candidate genes for CRC risk in pathways or gene families not previously implicated by GWAS. Locus 13q22.1, represented by lead SNP rs78341008 (MAF 7.2%; P=3.2×10^{-10}), is near *KLF5*, a known CRC oncogene that can be activated by somatic hotspot mutations or super-enhancer duplications[31,32]. *KLF5* encodes transcription factor Krüppel-like factor 5 (KLF5), which promotes cell proliferation and is highly expressed in intestinal crypt stem cells. We also found an association at 19p13.11, near *KLF2*. KLF2 expression in endothelial cells is critical for normal blood vessel function[33,34]. Down-regulated KLF2 expression in colon tumor tissues contributes to structurally and functionally abnormal tumor blood vessels, resulting in impaired blood flow and hypoxia in tumors[35]. Another locus at 9q31.1 is near *LPAR1*, which encodes a receptor for lysophosphatidic acid (LPA). LPA-induced expression of hypoxia-inducible factor 1 (HIF-1α), a key regulator of cellular adaptation to hypoxia and tumorigenesis, depends on KLF5[36]. Additionally, LPA activates multiple signaling pathways and stimulates proliferation of colon cancer cells by activation of *KLF5*[37]. Another locus (7p13) is near *SNHG15,* encoding a long non-coding RNA (lncRNA) that epigenetically represses *KLF2* to promote pancreatic cancer proliferation[38].

We found two loci near members of the Hedgehog (Hh) signaling pathway. Aberrant activation of this pathway, caused by somatic mutations or changes in expression, can drive tumorigenesis in many tumors[39]. Notably, downregulated stromal cell Hh signaling reportedly accelerates colonic tumorigenesis in mice[40]. Locus 3q13.2, represented by low-frequency lead SNP rs72942485 (MAF 2.2%; P=2.1×10^{-8}), overlaps *BOC*, encoding a Hh coreceptor molecule. In medulloblastoma, upregulated *BOC* promotes Hh-driven tumor progression through Cyclin D1-induced DNA damage[41]. In pancreatic cancer, a complex role for stromal *BOC* expression in tumorigenesis and angiogenesis has been reported[42]. Locus 4q31.21 is near *HHIP*, encoding an inhibitor of Hh signaling. Of note, the Hh signaling pathway was also significantly enriched in our pathway analysis (described below).

Locus 11q22.1 is near *YAP1*, which encodes a critical downstream regulatory target in the Hippo signaling pathway that is gaining recognition as a pivotal player in organ size control and tumorigenesis[43]. *YAP1* is highly expressed in intestinal crypt stem cells, and in transgenic mice, overexpression resulted in severe intestinal dysplasia and loss of differentiated cell types[44], reminiscent of phenotypes observed in mice and humans with deleterious germline *APC* mutations. Further, Hypoxia-inducible factor 2α (HIF-2α) promotes colon cancer growth by up-regulating YAP1 activity[45].

We provide further evidence for a link between immune function and CRC pathogenesis, and implicate the major histocompatibility complex (MHC) in CRC risk. We identified a locus near genes *HLA-DRB1*/*HLA-DQA1*, which is associated with immune-mediated diseases[46].

We identified two new loci near known tumor suppressor genes. Locus 4q24 is near *TET2*, a chromatin-remodeling gene frequently somatically mutated in multiple cancers, including colon cancer[47], and overlapping GWAS signals for multiple other cancers[48–50]. The *CDKN2B-CDKN2A-ANRIL* locus at 9p21.3 is a well-established hot spot of pleiotropic GWAS associations for many complex diseases including coronary artery disease[51], type 2 diabetes[52], and cancers[50,53,54–56]. Interestingly, lead variant rs1537372 is in high LD (r^2=0.82) with variants associated with coronary artery disease[51] and endometriosis[57], but not with the other cancer-associated variants. *CDKN2A/B* encode cyclin-dependent kinase inhibitors that regulate the cell

cycle. *CDKN2A* is one of the most commonly inactivated genes in cancer, and is a high penetrance gene for melanoma[58,59]. *CDKN2B* activation is tightly controlled by the cytokine TGF-β, further linking this signaling pathway with CRC tumorigenesis[60].

Our findings implicate genes in pathways with established roles in CRC pathogenesis. We identified loci at *SMAD3* and *SMAD9*, members of the TGF-β signaling pathway that includes genes linked to familial CRC syndromes (e.g., *SMAD4* and *BMPR1A*) and several GWAS-implicated genes (e.g., *SMAD7*, *BMP2*, *BMP4*)[61]. We identified another locus near TGF-β Receptor 1 (*TGFBR1*). Nearby gene *GALNT12* reportedly harbors inactivating germline and somatic mutations in human colon cancers[62] and, therefore, could also be the regulated effector gene. We identified a locus at 14q23.1 near *DACT1*, a member of the Wnt-β-catenin pathway with genes previously linked to familial CRC syndromes (*APC*[63]), and several GWAS-implicated genes (e.g., *CTNNB1*[18] and *TCF7L2*[17]). Genes related to telomere biology were linked by other GWAS: *TERC*[10] and *TERT*[22], encoding the RNA and protein subunit of telomerase respectively, and *FEN1*[17], involved in telomere stability[64]. A new locus at 20q13.33 harbors another gene related to telomere biology, *RTEL1*. This gene is involved in DNA double-strand break repair, and overlaps GWAS signals for cancers[55,65] and inflammation-related phenotypes, including inflammatory bowel disease[66] and atopic dermatitis[67].

Of 61 signals at 56 loci previously associated with CRC at GWS, 42 showed association evidence at $P < 5\times10^{-8}$ in the combined meta-analysis, and 55 at $P < 0.05$ in the independent Stage 2 meta-analysis (**Supplementary Table 3**). Of note, the association of rs755229494 at locus 5q22.2 (P=2.1×10^{-12}) was driven by studies with predominantly Ashkenazi Jewish ancestry and this SNP is in perfect LD with known missense SNP rs1801155 in the *APC* gene (I1307K), the minor allele of which is enriched in this population (MAF 6%), but rare in other populations[68,69].

Delineating distinct association signals at CRC risk loci

To identify additional independent association signals at known or new CRC risk loci, we conducted conditional analysis using individual-level data of 125,478 participants (Online Methods). At nine loci we observed 10 new independent association signals that attained P_J

$<5\times10^{-8}$ in a joint multiple-variant analysis (**Table 2**; **Supplementary Table 4**; **Supplementary Fig. 5**). Because this analysis focused on <5% of the genome, we also report signals at $P_J < 1\times10^{-5}$ in **Supplementary Table 5**. At 22 loci, we observed 25 new suggestive associations with $P_J < 1\times10^{-5}$.

At 11q13.4, near *POLD3* and *CHRDL2*, we identified a new low-frequency variant (lead SNP rs61389091, MAF 3.94%) separated by a recombination hotspot from the known common variant signal[12] (LD r^2 between lead SNPs <0.01). At 5p15.33, we identified another lower-frequency variant association (lead SNP rs78368589, MAF 5.97%), which was independent from the previously reported common variant signal 56kb away near *TERT* and *CLPTM1L* (LD r^2 with lead SNP rs2735940 <0.01)[22]. Variants in this region were linked to many cancer types, including lung, prostate, breast, and ovarian cancer[70].

The remaining eight new signals involved common variants. At new locus 2q33.1, near genes *PLCL1* and *SATB2*, two statistically independent associations (LD r^2 between two lead SNPs <0.01) are separated by a recombination hotspot (**Supplementary Fig. 5**). In the MHC region, we identified a conditionally independent signal near genes involved in NF-κβ signaling, including the gene encoding tumor necrosis factor-α, genes for the stress-signaling proteins MICA/MICB, and *HLA-B*. Locus 20p12.3, near *BMP2*, harbored four distinct association signals (**Figure 1**), two of which were reported previously[10,11] (**Supplementary Table 5**). All four SNPs selected in the model were in pairwise linkage equilibrium (maximum LD $r^2 = 0.039$, between rs189583 and rs994308). Our conditional analysis further confirmed that the signal ~1-Mb centromeric of *BMP2*, near gene *HAO1*, is independent. At 8q24.21 near *MYC*, the locus showing the second strongest statistical evidence of association in the combined meta-analysis (lead SNP rs6983267; $P = 3.4\times10^{-64}$), we identified a second independent signal (lead SNP rs4313119, $P_J = 2.1\times10^{-9}$; LD r^2 with rs6983267 <0.001). At the recently reported locus 5p13.1[22], near the non-coding RNA gene *LINC00603*, we identified an additional signal (lead SNP rs7708610) that was partly masked by the reported signal in the single-variant analysis due to the negative correlation between rs7708610 and rs12514517 ($r = -0.18$; $r^2 = 0.03$). This caused significance for both SNPs to increase markedly when fitted jointly (rs7708610, unconditional $P = 1.5\times10^{-5}$ and $P_J = 3.8\times10^{-9}$). At 12p13.32 near *CCND2*, we identified a new

signal (lead SNP rs3217874, $P_J = 2.4\times10^{-9}$) and confirmed two previously associated signals[13–15] (**Supplementary Text**). At the *GREM1* locus on 15q13.3, two independent signals were previously described[11]. Our analyses suggest that this locus harbors three signals. A new signal represented by SNP rs17816465 is conditionally independent from the other two signals ($P_J = 1.4\times10^{-10}$, conditioned on rs2293581 and rs12708491; LD with conditioning SNPs $r^2<0.01$; **Supplementary Text**).

Additionally, signals with P_J values approaching GWS were observed at new locus 3q13.2 near *BOC* (rs13086367, unconditional $P = 6.7\times10^{-8}$, $P_J = 6.9\times10^{-8}$, MAF=47.4%), 96kb from the low-frequency signal represented by rs72942485 (unconditional $P = 2.1\times10^{-8}$, $P_J = 1.3\times10^{-8}$, MAF=2.2%); at known locus 10q22.3 near *ZMIZ1* (rs1250567, unconditional $P = 3.1\times10^{-8}$, $P_J = 7.2\times10^{-8}$, MAF=45.1%); and at new locus 13q22.1 near *KLF5* (rs45597035, unconditional $P = 2.7\times10^{-9}$, $P_J = 8.1\times10^{-8}$, MAF=34.4%) (**Supplementary Table 5**). Furthermore, we clarify previously reported independent association signals (**Supplementary Text**).

Associations of CRC risk variants with other traits

Nineteen of the GWS association signals for CRC were in high LD ($r^2>0.7$) with at least one SNP in the NHGRI-EBI GWAS Catalog[46] that has significant association in GWAS of other traits. Notable overlap included SNPs associated with other cancers, immune-related traits (e.g., tonsillectomy, inflammatory bowel disease, and circulating white blood cell traits), obesity traits, blood pressure, and other cardiometabolic traits (**Supplementary Table 6**).

Mechanisms underlying CRC association signals

To further localize variants driving the 40 newly identified signals, we used association evidence to define credible sets of variants that are 99% likely to contain the causal variant (Online Methods). The 99% credible set size for new loci ranged from one (17p12) to 93 (2q33.1). For 11 distinct association signals, the set included ten or fewer variants (**Supplementary Table 7**). At locus 17p12, we narrowed the candidate variant to rs1078643, located in exon 1 of the lncRNA *LINC00675* that is primarily expressed in gastrointestinal tissues. Small credible sets were observed for locus 4q31.21 (two variants, indexed by synonymous SNP rs11727676 in *HHIP*), and signals at known loci near *GREM1* (one variant) and *CCND2* (two variants).

We performed functional annotation of credible set variants to nominate putative causal variants. Eight sets contained coding variants but only the synonymous SNP in *HHIP* had a high posterior probability of driving the association (**Supplementary Table 8**). Next, we examined overlap of credible sets with regulatory genomic annotations from 51 existing CRC-relevant datasets to examine non-coding functions (Online Methods). Also, to better refine regulatory elements in active enhancers, we performed ATAC-seq to measure chromatin accessibility in four colonic crypts and used resulting data to annotate GWAS signals.

Of the 40 sets, 36 overlapped with active enhancers identified by histone mark H3K27ac measured in normal colonic crypt epithelium, CRC cell lines, or CRC tissue (**Supplementary Table 8**; **Supplementary Fig. 6**). Twenty of these 36 overlapped with super-enhancers. Notably, when compared with epigenomics data from normal colonic crypt epithelium, all 36 sets overlapped enhancers with gained or lost activity in one or more CRC specimens. Eleven of these sets overlapped enhancers recurrently gained or lost in $\geq$20 CRC cell lines.

The locus at GWAS hot spot 9p21 overlaps a super-enhancer, and the credible set is entirely intronic to *ANRIL*, alias *CDKN2B-AS1*. The Genotype-Tissue Expression (GTEx) data show that the antisense lncRNA *ANRIL* is exclusively expressed in transverse colon and small intestine. Interestingly, ANRIL recruits SUZ12 and EHZ2 to epigenetically silence tumor suppressor genes *CDKN2A/B*[71].

Noncoding somatic driver mutations or focal amplifications have been reported in regions regulating expression of *MYC*[72], *TERT*[73], and *KLF5*[31], now implicated by GWAS for CRC. We checked whether GWAS-identified association signals co-localize with these regions and found that the *KLF5* signal overlaps the somatically amplified super-enhancer flanked by *KLF5* and *KLF12* (**Figure 2**). Also, the previously reported signal in the *TERT* promotor region[22] overlaps with the recurrent somatically mutated region in multiple cancers[73].

To test whether CRC associations are non-randomly distributed across genomic features, we used GARFIELD[74]. Focusing on DNase I hypersensitive site (DHS) peaks that identify open

chromatin, we observed significant enrichment across many cell types, particularly fetal tissues, with strongest enrichment observed in fetal gastrointestinal tissues, $CD20^+$ primary cells (B cells), and embryonic stem cells (**Supplementary Fig. 7**; **Supplementary Table 9)**.

We used MAGENTA[75] to identify pathways or gene sets enriched for associations with CRC, assessing two gene *P*-value cutoffs: 95th and 75th percentiles. At the 75th percentile, we observed enrichment of multiple KEGG cancer pathways at a false discovery rate (FDR) of 0.05. This was not observed for the 95th percentile cutoff and suggests that many more loci that are shared with other cancer types remain to be identified in larger studies. Using the 75th (95th) percentile cutoff, at FDR 0.05 and 0.20, we found enrichment of 7 (5) and 53 (24) gene sets, respectively. Established pathways related to TGF-β/SMAD and BMP signaling were among the top enriched pathways. Other notable enriched pathways included Hedgehog signaling, basal cell carcinoma, melanogenesis, cell cycle, S phase, and telomere maintenance (**Supplementary Table 10**).

Polygenicity of colorectal cancer and contribution of rare variants

To estimate the contribution of rare variants (MAF ≤1%) to CRC heritability, we used the LD- and MAF-stratified component GREML (GREML-LDMS) method implemented in GCTA[76] (Online Methods). Assuming a lifetime risk of 4.3%, we estimated that all imputed autosomal variants explain 21.6% (95% CI=17.5-25.7%) of the variation in liability for CRC, with almost half of this contributed by rare variants (h_g^2= 9.7%, 95% CI=6.2-13.3%; likelihood ratio test *P*=0.003); the estimated liability-scale heritability for variants with MAF >1% is 11.8% (95% CI=8.9-14.7%). Our overall estimate falls within the range of heritability reported by large twin studies[2]. Because heritability estimates for rare variants are sensitive to potential biases due to technical effects or population stratification[77] and the contribution of rare variants is probably underestimated due to limitations of genotype imputation, results should be interpreted with caution. Overall, findings suggest that missing heritability is not large, but that many rare and common variants have yet to be identified.

Familial relative risk explained by GWAS-identified variants

Adjusting for winner's curse[78], the familial relative risk (RR) to first-degree relatives (λ_0) attributable to GWAS-identified variants rose from 1.072 for the 55 previously described autosomal risk variants that showed evidence for replication at P <0.05, to 1.092 after inclusion of 40 new signals, and increased further to 1.098 when we included 25 suggestive association signals reported in **Supplementary Table 5** (Online Methods). Assuming a λ_0 of 2.2, the 55 established signals account for 8.8% of familial RR explained (95% CI: 8.1-9.4). Established signals combined with 40 newly discovered signals account for 11.2% (95% CI: 10.5-12.0), and adding 25 suggestive signals increases this to 11.9% (95% CI: 11.1-12.7).

Implications for stratified screening prevention

We demonstrate how using a polygenic risk score (PRS) derived from 95 independent association signals could impact clinical guidelines for preventive screening. The difference in recommended starting age for screening for those in the highest 1% (and 10%) percentiles of risk compared with lowest percentiles is 18 years (and 10 years) for men, and 24 years (and 12 years) for women (**Figure 3**; Online Methods). **Supplementary Table 11** gives risk allele frequency (RAF) estimates in different populations for variants included in the PRS. As expected, RAFs vary across populations. Furthermore, differences in LD between tagging and true causal variants across populations can result in less prediction accuracy and subsequent lower predictive power of the PRS in non-European populations. Accordingly, it will be important to develop ancestry-specific PRSs that incorporate detailed fine-mapping results for each GWAS signal.

DISCUSSION

To further define the genetic architecture of sporadic CRC, we performed low-coverage WGS and imputation into a large set of GWAS data. We discovered 40 new CRC signals and replicated 55 previously reported signals. We found the first rare variant signal for sporadic CRC, which represents the strongest protective rare allelic effect identified to date. Our analyses highlight new genes and pathways contributing to underlying CRC risk and suggest roles for Krüppel-like factors, Hedgehog signaling, Hippo-YAP signaling, and immune function. Multiple loci provide new evidence for an important role of lncRNAs in CRC tumorigenesis[79]. Functional genomic annotations support that most sporadic CRC genetic risk lies in non-coding genomic regions. We further show how newly discovered variants can lead to improved risk prediction.

This study underscores the critical importance of large-scale GWAS collaboration. While discovery of the rare variant signal was only possible through increased coverage and improved imputation accuracy enabled by imputation panels, sample size was pivotal for discovery of new CRC loci. Results suggest that CRC exhibits a highly polygenic architecture, much of which remains undefined. This also suggests that continued GWAS efforts, together with increasingly comprehensive imputation panels that allow for improved low-frequency and rare genetic variant imputation, will uncover more CRC risk variants. In addition, to investigate sites that are not imputable, large-scale deep sequencing will be needed. Importantly, the prevailing European bias in CRC GWAS limits the generalizability of findings and the application of PRSs in non-European (especially African) populations[80]. Therefore, a broader representation of ancestries in CRC GWAS is necessary.

Studies of somatic genomic alterations in cancer have mostly focused on the coding genome and identification of noncoding drivers has proven to be challenging[73]. Yet, noncoding somatic driver mutations or focal amplications in regulatory regions impacting expression have been reported for *MYC*[72], *TERT*[73], and *KLF5*[31]. The observed overlap between GWAS-identified CRC risk loci and somatic driver regions strongly suggests that expanding the search of somatic driver mutations to noncoding regulatory elements will yield additional discoveries and that searches for somatic drivers can be guided by GWAS findings.

Additionally, we found loci near proposed drug targets, including *CHD1*, implicated by the rare variant signal, and *KLF5*. To date, cancer drug target discovery research has almost exclusively focused on properties of cancer cells, yielding drugs that target proteins either highly expressed or expressed in a mutant form due to frequent recurrent somatic missense mutations (e.g., $BRAF^{V600E}$) or gene fusion events. In stark contrast with other common complex diseases, cancer GWAS results are not being used extensively to inform drug target selection. It has been estimated that selecting targets supported by GWAS could double the success rate in clinical development[81]. Our discoveries corroborate that not using GWAS results to inform drug discovery is a missed opportunity, not only for treating cancers, but also for chemoprevention in high-risk individuals.

In summary, in the largest genome-wide scan for sporadic CRC risk thus far, we identified the first rare variant signal for sporadic CRC, and almost doubled the number of known association signals. Our findings provide a substantial number of new leads that may spur downstream investigation into the biology of CRC risk, and that will impact drug development and clinical guidelines, such as personalized screening decisions.

Acknowledgements

A full list of acknowledgements appears in the Supplementary Notes.

Author contributions

J.R.H., S.A.B. and T.A.H. contributed equally, and D.A.N., S.B.G., L.H. and U.P. jointly supervised this research. J.R.H., S.A.B., T.A.H., H.M.K., D.V.C., M.W., F.R.S., J.D.S., D.A., M.H.A., K.A., C.A.-C., V.A., C.B., J.A.B., S.I.B., S.B., D.T.B., J.B., H. Boeing, H. Brenner, S. Brezina, S. Buch, D.D.B., A.B.-H., K.B., B.J.C., P.T.C., S.C.-B., A.T.C., J.C.-C., S.J.C., M.-D.C., S.H.C., A.J.C., K.C., A.d.l.C., D.F.E., S.G.E., F.E., D.R.E., E.J.M.F., J.C.F., D.F., S.G., G.G.G., E.G., P.J.G., J.S.G., A.G., M.J.G., R.W.H., J.H., H.H., R.B.H., P.H., M.H., J.L.H., W.-Y.H., T.J.H., D.J.H., R.J., E.J.J., M.A.J., T.O.K., T.J.K., H.R.K., L.N.K., C.K., S.K., S.-S.K., L.L.M., S.C.L., C.I.L., L.L., A.L., N.M.L., S.M., S.D.M., V.M., G.M., M.M., R.L.M., L.M., R.M., A.N., P.A.N., K.O., N.C.O.-M., B.P., P.S.P., R.P., V.P., P.D.P.P., E.A.P., R.L.P., G.R., H.S.R., E.R., M.R.-B., C.S., R.E.S., D.S., M.-H.S., S.S., M.L.S., C.M.T., S.N.T., A.T., C.M.U., F.J.B.v.D., B.V.G., H.v.K., J.V., K.V., P.V., L.V., V.V., E.W., C.R.W., A.W., M.O.W., A.H.W., B.W.Z., W.Z., P.C.S., J.D.P., M.C.B., G.C., V.M., G.R.A., D.A.N., S.B.G., L.H. and U.P. conceived and designed the experiments. T.A.H., M.W., J.D.S., K.F.D., D.D., R.I., E.K., H.L., C.E.M., E.P., J.R., T.S., S.S.T., D.J.V.D.B., M.C.B., D.A.N. performed the experiments. J.R.H., H.M.K.., S.C., S.L.S., D.V.C., C.Q., J.J., C.K.E., P.G., F.R.S., D.M.L., S.C.N., N.A.S.-A., C.A.L., M.L., T.L.L., Y.-R.S., A.K., G.R.A., L.H. performed statistical analysis. J.R.H., S.A.B., T.A.H., H.M.K.., S.C., S.L.S., D.V.C., C.Q., J.J., C.K.E., P.G., M.W., F.R.S., D.M.L., S.C.N., N.A.S.-A., B.L.B., C.S.C., C.M.C., K.R.C., J.G., W.-L.H., C.A.L., S.M.L., M.L., Y.L., T.L.L., M.S., Y.-R.S., A.K., G.R.A., L.H., U.P. analyzed the data. H.M.K., C.K.E., D.A., M.H.A., K.A., C.A.-C., V.A., C.B., J.A.B., S.I.B., S.B., D.T.B., J.B., H. Boeing, H. Brenner, S. Brezina, S.

Buch, D.D.B., A.B.-H., K.B., B.J.C., P.T.C., S.C.-B., A.T.C., J.C.-C., S.J.C., M.-D.C., S.H.C., A.J.C., K.C., A.d.l.C., D.F.E., S.G.E., F.E., D.R.E., E.J.M.F., J.C.F., R.F., L.M.F., D.F., M.G., S.G., W.J.G., G.G.G., P.J.G., W.M.G., J.S.G., A.G., M.J.G., R.W.H., J.H., H.H., S.H., R.B.H., P.H., M.H., J.L.H., W.-Y.H., T.J.H., D.J.H., G.I.-S., G.E.I., R.J., E.J.J., M.A.J., A.D.J., C.E.J., T.O.K., T.J.K., H.R.K., L.N.K., C.K., T.K., S.K., S.-S.K., S.C.L., L.L.M., S.C.L., F.L., C.I.L., L.L., W.L., A.L., N.M.L., S.M., S.D.M., V.M., G.M., M.M., R.L.M., L.M., N.M., R.M., A.N., P.A.N., K.O., S.O, N.C.O.-M., B.P., P.S.P., R.P., V.P., P.D.P.P., M.P., E.A.P., R.L.P., L.R., G.R., H.S.R., E.R., M.R.-B., L.C.S., C.S., R.E.S., M.S., M.-H.S., K.S., S.S., M.L.S., M.C.S., Z.K.S., C.S., C.M.T., S.N.T., D.C.T., A.E.T., A.T., C.M.U., F.J.B.v.D., B.V.G., H.v.K., J.V., K.V., P.V., L.V., V.V., K.W., S.J.W., E.W., A.K.W., C.R.W., A.W., M.O.W., A.H.W., S.H.Z., B.W.Z., Q.Z., W.Z., P.C.S., J.D.P., M.C.B., A.K., G.C., V.M., G.R.A., S.B.G. and U.P. contributed reagents/materials/analysis tools. J.R.H., S.A.B., T.A.H., J.J., L.H. and U.P. wrote the paper.

Competing Interests Statement

Goncalo R Abecasis has received compensation from 23andMe and Helix. He is currently an employee of Regeneron Pharmaceuticals. Heather Hampel performs collaborative research with Ambry Genetics, InVitae Genetics, and Myriad Genetic Laboratories, Inc., is on the scientific advisory board for InVitae Genetics and Genome Medical, and has stock in Genome Medical. Rachel Pearlman has participated in collaborative funded research with Myriad Genetics Laboratories and Invitae Genetics but has no financial competitive interest.

REFERENCES

1. Ferlay, J. *et al.* Cancer incidence and mortality worldwide: sources, methods and major patterns in GLOBOCAN 2012. *Int J Cancer* **136,** E359-86 (2015).
2. Lichtenstein, P. *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* **343,** 78–85 (2000).
3. Czene, K., Lichtenstein, P. & Hemminki, K. Environmental and heritable causes of cancer among 9.6 million individuals in the Swedish Family-Cancer Database. *Int J Cancer* **99,** 260–266 (2002).
4. Sud, A., Kinnersley, B. & Houlston, R. S. Genome-wide association studies of cancer: current insights and future perspectives. *Nat Rev Cancer* **17,** 692–704 (2017).
5. Tomlinson, I. *et al.* A genome-wide association scan of tag SNPs identifies a susceptibility

variant for colorectal cancer at 8q24.21. *Nat Genet* **39,** 984–988 (2007).
6. Broderick, P. *et al.* A genome-wide association study shows that common alleles of SMAD7 influence colorectal cancer risk. *Nat Genet* **39,** 1315–1317 (2007).
7. Tomlinson, I. P. M. *et al.* A genome-wide association study identifies colorectal cancer susceptibility loci on chromosomes 10p14 and 8q23.3. *Nat Genet* **40,** 623–630 (2008).
8. Tenesa, A. *et al.* Genome-wide association scan identifies a colorectal cancer susceptibility locus on 11q23 and replicates risk loci at 8q24 and 18q21. *Nat Genet* **40,** 631–637 (2008).
9. COGENT Study *et al.* Meta-analysis of genome-wide association data identifies four new susceptibility loci for colorectal cancer. *Nat Genet* **40,** 1426–1435 (2008).
10. Houlston, R. S. *et al.* Meta-analysis of three genome-wide association studies identifies susceptibility loci for colorectal cancer at 1q41, 3q26.2, 12q13.13 and 20q13.33. *Nat Genet* **42,** 973–977 (2010).
11. Tomlinson, I. P. M. *et al.* Multiple common susceptibility variants near BMP pathway loci GREM1, BMP4, and BMP2 explain part of the missing heritability of colorectal cancer. *PLoS Genet* **7,** e1002105 (2011).
12. Dunlop, M. G. *et al.* Common variation near CDKN1A, POLD3 and SHROOM2 influences colorectal cancer risk. *Nat Genet* **44,** 770–776 (2012).
13. Peters, U. *et al.* Identification of Genetic Susceptibility Loci for Colorectal Tumors in a Genome-Wide Meta-analysis. *Gastroenterology* **144,** 799–807.e24 (2013).
14. Jia, W.-H. *et al.* Genome-wide association analyses in East Asians identify new susceptibility loci for colorectal cancer. *Nat Genet* **45,** 191–196 (2013).
15. Whiffin, N. *et al.* Identification of susceptibility loci for colorectal cancer in a genome-wide meta-analysis. *Hum Mol Genet* **23,** 4729–4737 (2014).
16. Wang, H. *et al.* Trans-ethnic genome-wide association study of colorectal cancer identifies a new susceptibility locus in VTI1A. *Nat Commun* **5,** 4613 (2014).
17. Zhang, B. *et al.* Large-scale genetic study in East Asians identifies six new loci associated with colorectal cancer risk. *Nat Genet* **46,** 533–542 (2014).
18. Schumacher, F. R. *et al.* Genome-wide association study of colorectal cancer identifies six new susceptibility loci. *Nat Commun* **6,** 7138 (2015).
19. Al-Tassan, N. A. *et al.* A new GWAS and meta-analysis with 1000Genomes imputation identifies novel risk variants for colorectal cancer. *Sci Rep* **5,** 10442 (2015).
20. Orlando, G. *et al.* Variation at 2q35 (PNKD and TMBIM1) influences colorectal cancer risk and identifies a pleiotropic effect with inflammatory bowel disease. *Hum Mol Genet* **25,** 2349–2359 (2016).
21. Zeng, C. *et al.* Identification of susceptibility loci and genes for colorectal cancer risk. *Gastroenterology* **150,** 1633–1645 (2016).
22. Schmit, S. L. *et al.* Novel common genetic susceptibility loci for colorectal cancer. *J Natl Cancer Inst* 1–12 (2018). doi:10.1093/jnci/djy099
23. Fuchsberger, C. *et al.* The genetic architecture of type 2 diabetes. *Nature* **536,** 41–47 (2016).
24. 1000 Genomes Project Consortium *et al.* A global reference for human genetic variation. *Nature* **526,** 68–74 (2015).
25. McCarthy, S. *et al.* A reference panel of 64,976 haplotypes for genotype imputation. *Nat Genet* **48,** 1279–1283 (2016).
26. Amos, C. I. *et al.* The oncoarray consortium: A network for understanding the genetic architecture of common cancers. *Cancer Epidemiol Biomarkers Prev* **26,** 126–135 (2017).

27. Zhao, D. & DePinho, R. A. Synthetic essentiality: Targeting tumor suppressor deficiencies in cancer. *Bioessays* **39,** (2017).
28. Zhao, D. *et al.* Synthetic essentiality of chromatin remodelling factor CHD1 in PTEN-deficient cancer. *Nature* **542,** 484–488 (2017).
29. Xiao, Y. *et al.* RGMb is a novel binding partner for PD-L2 and its engagement with PD-L2 promotes respiratory tolerance. *J Exp Med* **211,** 943–959 (2014).
30. Topalian, S. L. *et al.* Safety, activity, and immune correlates of anti-PD-1 antibody in cancer. *N Engl J Med* **366,** 2443–2454 (2012).
31. Zhang, X. *et al.* Somatic superenhancer duplications and hotspot mutations lead to oncogenic activation of the KLF5 transcription factor. *Cancer Discov* **8,** 108–125 (2018).
32. Giannakis, M. *et al.* Genomic Correlates of Immune-Cell Infiltrates in Colorectal Carcinoma. *Cell Rep* **15,** 857–865 (2016).
33. Dekker, R. J. *et al.* KLF2 provokes a gene expression pattern that establishes functional quiescent differentiation of the endothelium. *Blood* **107,** 4354–4363 (2006).
34. Boon, R. A. *et al.* KLF2 suppresses TGF-beta signaling in endothelium through induction of Smad7 and inhibition of AP-1. *Arterioscler Thromb Vasc Biol* **27,** 532–539 (2007).
35. Chakroborty, D. *et al.* Dopamine stabilizes tumor blood vessels by up-regulating angiopoietin 1 expression in pericytes and Kruppel-like factor-2 expression in tumor endothelial cells. *Proc Natl Acad Sci U S A* **108,** 20730–20735 (2011).
36. Lee, S.-J. *et al.* Regulation of hypoxia-inducible factor 1α (HIF-1α) by lysophosphatidic acid is dependent on interplay between p53 and Krüppel-like factor 5. *J Biol Chem* **288,** 25244–25253 (2013).
37. Zhang, H. *et al.* Lysophosphatidic acid facilitates proliferation of colon cancer cells via induction of Krüppel-like factor 5. *J Biol Chem* **282,** 15541–15549 (2007).
38. Ma, Z. *et al.* Long non-coding RNA SNHG15 inhibits P15 and KLF2 expression to promote pancreatic cancer proliferation through EZH2-mediated H3K27me3. *Oncotarget* **8,** 84153–84167 (2017).
39. Evangelista, M., Tian, H. & de Sauvage, F. J. The hedgehog signaling pathway in cancer. *Clin Cancer Res* **12,** 5924–5928 (2006).
40. Gerling, M. *et al.* Stromal Hedgehog signalling is downregulated in colon cancer and its restoration restrains tumour growth. *Nat Commun* **7,** 12321 (2016).
41. Mille, F. *et al.* The Shh receptor Boc promotes progression of early medulloblastoma to advanced tumors. *Dev Cell* **31,** 34–47 (2014).
42. Mathew, E. *et al.* Dosage-dependent regulation of pancreatic cancer growth and angiogenesis by hedgehog signaling. *Cell Rep* **9,** 484–494 (2014).
43. Zhao, B., Li, L., Lei, Q. & Guan, K.-L. The Hippo-YAP pathway in organ size control and tumorigenesis: an updated version. *Genes Dev* **24,** 862–874 (2010).
44. Camargo, F. D. *et al.* YAP1 increases organ size and expands undifferentiated progenitor cells. *Curr Biol* **17,** 2054–2060 (2007).
45. Ma, X., Zhang, H., Xue, X. & Shah, Y. M. Hypoxia-inducible factor 2α (HIF-2α) promotes colon cancer growth by potentiating Yes-associated protein 1 (YAP1) activity. *J Biol Chem* **292,** 17046–17056 (2017).
46. MacArthur, J. *et al.* The new NHGRI-EBI Catalog of published genome-wide association studies (GWAS Catalog). *Nucleic Acids Res* **45,** D896–D901 (2017).
47. Seshagiri, S. *et al.* Recurrent R-spondin fusions in colon cancer. *Nature* **488,** 660–664 (2012).

48. Song, F. *et al.* Identification of a melanoma susceptibility locus and somatic mutation in TET2. *Carcinogenesis* **35,** 2097–2101 (2014).
49. Eeles, R. A. *et al.* Identification of seven new prostate cancer susceptibility loci through a genome-wide association study. *Nat Genet* **41,** 1116–1121 (2009).
50. Michailidou, K. *et al.* Association analysis identifies 65 new breast cancer risk loci. *Nature* **551,** 92–94 (2017).
51. Schunkert, H. *et al.* Large-scale association analysis identifies 13 new susceptibility loci for coronary artery disease. *Nat Genet* **43,** 333–338 (2011).
52. Scott, L. J. *et al.* A genome-wide association study of type 2 diabetes in Finns detects multiple susceptibility variants. *Science* **316,** 1341–1345 (2007).
53. Al Olama, A. A. *et al.* A meta-analysis of 87,040 individuals identifies 23 new susceptibility loci for prostate cancer. *Nat Genet* **46,** 1103–1109 (2014).
54. Timofeeva, M. N. *et al.* Influence of common genetic variation on lung cancer risk: meta-analysis of 14 900 cases and 29 485 controls. *Hum Mol Genet* **21,** 4980–4995 (2012).
55. Shete, S. *et al.* Genome-wide association study identifies five susceptibility loci for glioma. *Nat Genet* **41,** 899–904 (2009).
56. Bishop, D. T. *et al.* Genome-wide association study identifies three loci associated with melanoma risk. *Nat Genet* **41,** 920–925 (2009).
57. Sapkota, Y. *et al.* Meta-analysis identifies five novel loci associated with endometriosis highlighting key genes involved in hormone metabolism. *Nat Commun* **8,** 15539 (2017).
58. Cannon-Albright, L. A. *et al.* Assignment of a locus for familial melanoma, MLM, to chromosome 9p13-p22. *Science* **258,** 1148–1152 (1992).
59. Hussussian, C. J. *et al.* Germline p16 mutations in familial melanoma. *Nat Genet* **8,** 15–21 (1994).
60. Seoane, J. *et al.* TGFbeta influences Myc, Miz-1 and Smad to control the CDK inhibitor p15INK4b. *Nat Cell Biol* **3,** 400–408 (2001).
61. Jung, B., Staudacher, J. J. & Beauchamp, D. Transforming Growth Factor β Superfamily Signaling in Development of Colorectal Cancer. *Gastroenterology* **152,** 36–52 (2017).
62. Guda, K. *et al.* Inactivating germ-line and somatic mutations in polypeptide N-acetylgalactosaminyltransferase 12 in human colon cancers. *Proc Natl Acad Sci U S A* **106,** 12921–12925 (2009).
63. Groden, J. *et al.* Identification and characterization of the familial adenomatous polyposis coli gene. *Cell* **66,** 589–600 (1991).
64. Saharia, A. *et al.* FEN1 ensures telomere stability by facilitating replication fork re-initiation. *J Biol Chem* **285,** 27057–27066 (2010).
65. Eeles, R. A. *et al.* Identification of 23 new prostate cancer susceptibility loci using the iCOGS custom genotyping array. *Nat Genet* **45,** 385–91, 391e1 (2013).
66. Liu, J. Z. *et al.* Association analyses identify 38 susceptibility loci for inflammatory bowel disease and highlight shared genetic risk across populations. *Nat Genet* **47,** 979–986 (2015).
67. Paternoster, L. *et al.* Multi-ancestry genome-wide association study of 21,000 cases and 95,000 controls identifies new risk loci for atopic dermatitis. *Nat Genet* **47,** 1449–1456 (2015).
68. Laken, S. J. *et al.* Familial colorectal cancer in Ashkenazim due to a hypermutable tract in APC. *Nat Genet* **17,** 79–83 (1997).
69. Niell, B. L., Long, J. C., Rennert, G. & Gruber, S. B. Genetic anthropology of the

colorectal cancer-susceptibility allele APC I1307K: evidence of genetic drift within the Ashkenazim. *Am J Hum Genet* **73,** 1250–1260 (2003).
70. Karami, S. *et al.* Telomere structure and maintenance gene variants and risk of five cancer types. *Int J Cancer* **139,** 2655–2670 (2016).
71. Congrains, A., Kamide, K., Ohishi, M. & Rakugi, H. ANRIL: molecular mechanisms and implications in human health. *Int J Mol Sci* **14,** 1278–1292 (2013).
72. Zhang, X. *et al.* Identification of focally amplified lineage-specific super-enhancers in human epithelial cancers. *Nat Genet* **48,** 176–182 (2016).
73. Rheinbay, E. *et al.* Discovery and characterization of coding and non-coding driver mutations in more than 2,500 whole cancer genomes. *BioRxiv* (2017). doi:10.1101/237313
74. Iotchkova, V. *et al.* GARFIELD - GWAS Analysis of Regulatory or Functional Information Enrichment with LD correction. *BioRxiv* (2016). doi:10.1101/085738
75. Segrè, A. V. *et al.* Common inherited variation in mitochondrial genes is not enriched for associations with type 2 diabetes or related glycemic traits. *PLoS Genet* **6,** (2010).
76. Yang, J. *et al.* Genetic variance estimation with imputed variants finds negligible missing heritability for human height and body mass index. *Nat Genet* **47,** 1114–1120 (2015).
77. Bhatia, G. *et al.* Subtle stratification confounds estimates of heritability from rare variants. *BioRxiv* (2016). doi:10.1101/048181
78. Zhong, H. & Prentice, R. L. Bias-reduced estimators and confidence intervals for odds ratios in genome-wide association studies. *Biostatistics* **9,** 621–634 (2008).
79. Cheetham, S. W., Gruhl, F., Mattick, J. S. & Dinger, M. E. Long noncoding RNAs and the genetics of cancer. *Br J Cancer* **108,** 2419–2425 (2013).
80. Popejoy, A. B. & Fullerton, S. M. Genomics is failing on diversity. *Nature* **538,** 161–164 (2016).
81. Nelson, M. R. *et al.* The support of human genetic evidence for approved drug indications. *Nat Genet* **47,** 856–860 (2015).

FIGURE LEGENDS

Figure 1 Conditionally independent association signals at the *BMP2* locus. Regional association plot showing the unconditional $-\log_{10}$(*P*-value) for the association with CRC risk in the combined meta-analysis of up to 125,478 individuals, as a function of genomic position (Build 37) for each variant in the region. The lead variants are indicated by a diamond symbol and its positions are indicated by dashed vertical lines. The color-labeling and shape of all other variants indicate the lead variant with which they are in strongest LD. The two new genome-wide significant signals are indicated by an asterisk.

Figure 2 Functional genomic annotation of new CRC risk locus overlapping *KLF5* super-enhancer. Top: Regional association plot showing the unconditional $-\log_{10}$(*P*-value) for the association with CRC risk in the combined meta-analysis of up to 125,478 individuals, as a function of genomic position (Build 37) for each variant in the region. The lead variants are indicated by a diamond symbol and its positions are indicated by dashed vertical lines. The color-labeling and shape of all other variants indicate the lead variant with which they are in strongest LD. **Bottom:** UCSC genome browser annotations for region overlapping the super-enhancer flanked by *KLF5* and *KLF12*, and spanning variants in LD with rs78341008, and with two conditionally independent association signals indexed by rs45597035 and rs1924816. The region is annotated with the following tracks (from top to bottom): UCSC gene annotations; epigenomic profiles showing MACS2 peak calls as transparent overlays for different samples taken from non-diseased colonic crypt cells or colon tissue (purple) and from different primary CRC cell lines or tumor samples (teal); position of the lead variants and variants in LD with the lead; variants in the 99% credible set; the union of super-enhancers called using the ROSE package; gray bars highlight the targeted enhancers (e1,e3, and e4) previously shown by Zhang *et al.*[31] to have combinatorial effects on *KLF5* expression. ATAC-seq data newly generated for this study show high resolution annotation of putative binding regions within the active super-enhancer further fine-mapping putative causal variants at each of the three signals.

Figure 3 Recommended age to start CRC screening based on a polygenic risk score (PRS). The PRS was constructed using the 95 known and newly discovered variants. The horizontal lines represent the recommended age for the first endoscopy for an average-risk person in the

914 current screening guideline for CRC. The risk threshold to determine the age for the first
915 screening was set as the average of 10-year CRC risks for a 50-year-old man (1.25%) and
916 woman (0.68%), i.e. (1.25%+0.68%)/2 = 0.97%, who have not previously received an
917 endoscopy. Details are given in the Online Methods.
918

9 **Table 1 New CRC risk loci reaching genome-wide significance ($P < 5\times10^{-8}$) in the combined (Stage 1 and Stage 2) meta-analysis.**

							Stage 1 meta-analysis: up to 34,869 cases and 29,051 controls			Stage 2 meta-analysis: up to 23,262 cases and 38,296 controls			Combined meta-analysis: up to 58,131 cases and 67,347 controls		
Locus	**Nearby gene(s)**	**rsID lead variant**	**Chr.**	**Position (Build 37)**	**Alleles (risk/other)**	**RAF (%)**	**OR**	**95% CI**	***P***	**OR**	**95% CI**	***P***	**OR**	**95% CI**	***P***
Rare variants															
5q21.1	*RGMB*; *CHD1*	rs145364999*	5	98,206,082	T/A	99.69	1.57	1.20-2.05	9.0×10^{-4}	1.93	1.48-2.52	1.0×10^{-6}	1.74	1.45-2.10	6.3×10^{-9}
Low-frequency variants															
3q13.2	*BOC*	rs72942485	3	112,999,560	G/A	98.02	1.16	1.07-1.26	2.5×10^{-4}	1.23	1.12-1.35	1.5×10^{-5}	1.19	1.12-1.26	2.1×10^{-8}
Common variants															
1p34.3	*FHL3*	rs4360494§	1	38,455,891	G/C	45.39	1.05	1.03-1.08	2.9×10^{-5}	1.06	1.03-1.08	3.3×10^{-5}	1.05	1.04-1.07	3.8×10^{-9}
1p32.3	*TTC22*; *PCSK9*	rs12144319*	1	55,246,035	C/T	25.48	1.07	1.04-1.10	1.4×10^{-6}	1.07	1.04-1.10	5.5×10^{-6}	1.07	1.05-1.09	3.3×10^{-11}
2q24.2	*MARCH7; TANC1*	rs448513§	2	159,964,552	C/T	32.60	1.06	1.03-1.08	1.9×10^{-5}	1.05	1.02-1.08	5.8×10^{-4}	1.05	1.03-1.07	4.4×10^{-8}
2q33.1	*SATB2*	rs983402*	2	199,781,586	T/C	33.12	1.05	1.03-1.08	7.2×10^{-5}	1.08	1.05-1.11	1.0×10^{-8}	1.07	1.05-1.09	7.7×10^{-12}
3q22.2	*SLCO2A1*	rs10049390§	3	133,701,119	A/G	73.53	1.06	1.03-1.09	4.9×10^{-5}	1.07	1.04-1.10	1.8×10^{-5}	1.06	1.04-1.08	3.8×10^{-9}
4q24	*TET2*	rs1391441	4	106,128,760	A/G	67.20	1.05	1.02-1.07	1.5×10^{-4}	1.06	1.03-1.09	2.3×10^{-5}	1.05	1.03-1.07	1.6×10^{-8}
4q31.21	*HHIP*	rs11727676	4	145,659,064	C/T	9.80	1.08	1.03-1.13	4.5×10^{-4}	1.10	1.05-1.14	1.5×10^{-5}	1.09	1.06-1.12	2.9×10^{-8}
6p21.32	*HLA-DRB1*; *HLA-DQA1*	rs9271695*	6	32,593,080	G/A	79.54	1.09	1.06-1.13	1.3×10^{-7}	1.09	1.05-1.12	1.7×10^{-7}	1.09	1.07-1.12	1.1×10^{-13}
7p13	*MYO1G*; *SNHG15*; *CCM2*; *TBRG4*	rs12672022§	7	45,136,423	T/C	83.45	1.07	1.04-1.11	1.6×10^{-5}	1.06	1.03-1.10	4.4×10^{-4}	1.07	1.04-1.09	2.8×10^{-8}
9p21.3	*ANRIL*; *CDKN2A*; *CDKN2B*	rs1537372§	9	22,103,183	G/T	56.92	1.05	1.02-1.07	1.4×10^{-4}	1.06	1.03-1.08	2.4×10^{-5}	1.05	1.03-1.07	1.4×10^{-8}
9q22.33	*GALNT12*; *TGFBR1*	rs34405347§	9	101,679,752	T/G	90.34	1.08	1.04-1.13	5.5×10^{-5}	1.09	1.04-1.13	1.5×10^{-4}	1.09	1.05-1.12	3.1×10^{-8}
9q31.3	*LPAR1*	rs10980628	9	113,671,403	C/T	21.06	1.05	1.02-1.09	3.1×10^{-4}	1.08	1.05-1.11	1.3×10^{-6}	1.07	1.04-1.09	2.8×10^{-9}
11q22.1	*YAP1*	rs2186607	11	101,656,397	T/A	51.78	1.05	1.03-1.08	1.1×10^{-5}	1.05	1.03-1.08	3.3×10^{-5}	1.05	1.04-1.07	1.5×10^{-9}
12q12	*PRICKLE1*; *YAF2*	rs11610543§	12	43,134,191	G/A	50.13	1.05	1.03-1.08	1.1×10^{-5}	1.06	1.03-1.08	2.8×10^{-5}	1.05	1.04-1.07	1.3×10^{-9}
12q13.3	*STAT6*; *LRP1*; *NAB2*	rs4759277	12	57,533,690	A/C	35.46	1.07	1.04-1.09	8.4×10^{-7}	1.04	1.02-1.07	1.6×10^{-3}	1.05	1.04-1.07	9.4×10^{-9}
13q13.3	*SMAD9*	rs7333607*	13	37,462,010	G/A	23.50	1.09	1.06-1.12	2.5×10^{-8}	1.07	1.04-1.10	4.4×10^{-6}	1.08	1.06-1.10	6.3×10^{-13}
13q22.1	*KLF5*	rs78341008§	13	73,791,554	C/T	7.19	1.13	1.07-1.18	1.4×10^{-6}	1.11	1.05-1.16	4.8×10^{-5}	1.12	1.08-1.16	3.2×10^{-10}
13q34	*COL4A2*; *COL4A1*; *RAB20*	rs8000189	13	111,075,881	T/C	64.01	1.05	1.02-1.07	2.1×10^{-4}	1.07	1.04-1.10	1.3×10^{-6}	1.06	1.04-1.08	1.8×10^{-9}
14q23.1	*DACT1*	rs17094983§	14	59,189,361	G/A	87.73	1.10	1.07-1.15	8.4×10^{-8}	1.08	1.04-1.12	9.0×10^{-5}	1.09	1.06-1.12	4.6×10^{-11}

15q22.33	*SMAD3*	rs56324967*	15	67,402,824	C/T	67.57	1.07	1.04-1.10	2.2×10^{-7}	1.08	1.05-1.11	9.8×10^{-8}	1.07	1.05-1.09	1.1×10^{-13}
16q23.2	*MAF*	rs9930005§	16	80,043,258	C/A	43.03	1.05	1.03-1.08	1.3×10^{-5}	1.05	1.02-1.07	4.0×10^{-4}	1.05	1.03-1.07	2.1×10^{-8}
17p12	*LINC00675*	rs1078643*	17	10,707,241	A/G	76.36	1.07	1.04-1.10	9.2×10^{-6}	1.09	1.05-1.12	1.1×10^{-7}	1.08	1.05-1.10	6.6×10^{-12}
17q24.3	*LINC00673*	rs983318§	17	70,413,253	A/G	25.26	1.07	1.04-1.10	1.2×10^{-6}	1.05	1.02-1.08	8.0×10^{-4}	1.06	1.04-1.08	5.6×10^{-9}
17q25.3	*RAB40B*; *METRLN*	rs75954926*	17	81,061,048	G/A	65.68	1.10	1.07-1.13	9.4×10^{-11}	1.09	1.06-1.12	4.8×10^{-9}	1.09	1.07-1.11	3.0×10^{-18}
19p13.11	*KLF2*	rs34797592§	19	16,417,198	T/C	11.82	1.09	1.05-1.13	8.2×10^{-6}	1.09	1.05-1.13	1.2×10^{-5}	1.09	1.06-1.12	4.2×10^{-10}
19q13.43	*TRIM28*	rs73068325	19	59,079,096	T/C	18.26	1.06	1.03-1.09	2.1×10^{-4}	1.07	1.04-1.11	5.0×10^{-5}	1.07	1.04-1.09	4.2×10^{-8}
20q13.12	*TOX2*; *HNF4A*	rs6031311§	20	42,666,475	T/C	75.91	1.07	1.04-1.10	1.7×10^{-6}	1.05	1.02-1.08	7.6×10^{-4}	1.06	1.04-1.08	6.8×10^{-9}
20q13.33	*TNFRSF6B*; *RTEL1*	rs2738783§,¶	20	62,308,612	T/G	20.29	1.07	1.04-1.10	2.6×10^{-6}	1.05	1.02-1.08	3.3×10^{-3}	1.06	1.04-1.08	5.3×10^{-8}

Lead variant is the most associated variant at the locus. rsIDs based on NCBI dbSNP Build 150. Alleles are on the + strand. Chr.: Chromosome. RAF: Risk allele frequency, based on stage 2 data. OR, odds ratio estimate for the risk allele. All *P*-values reported in this table are based on fixed-effects inverse variance-weighted meta-analysis.

*Indicates that variant or LD proxy (r^2>0.7) was selected for our custom genotyping panel and formally replicates in the Stage 2 meta-analysis at a Bonferroni significance threshold of $P < 7.8\times10^{-6}$.

§Indicates that variant or LD proxy (r^2>0.7) was selected for our custom genotyping panel but did not attain Bonferroni significance in the Stage 2 meta-analysis.

¶This SNP reached genome-wide significance in the combined (Stage 1 + Stage 2) sample-size weighted meta-analysis based on likelihood ratio test results ($P = 4.9\times10^{-8}$).

Table 2 Additional new conditionally independent association signals at known and newly identified CRC risk loci that reach genome-wide significance ($P < 5\times10^{-8}$) in the combined meta-analysis of up to 125,478 individuals.

Locus	Nearby gene(s)	rsID lead variant	Chr.	Position (Build 37)	Alleles (risk/other)	RAF (%)	$OR_{unconditional}$	95% CI	$P_{unconditional}$	Joint multiple-variant analysis			
										Conditioning variant(s)	$OR_{conditional}$	95% CI	$P_{conditional}$
Low-frequency variants													
11q13.4	*POLD3*	rs61389091	11	74,427,921	C/T	96.06	1.23	1.18-1.29	1.2×10^{-18}	rs7121958*, rs7946853	1.21	1.16-1.27	3.7×10^{-16}
Common variants													
2q33.1	*SATB2*	rs11884596	2	199,612,407	C/T	38.23	1.06	1.04-1.08	1.1×10^{-9}	rs983402	1.06	1.04-1.07	3.6×10^{-9}
5p15.33	*TERT*; *CLPTM1L*	rs78368589	5	1,240,204	T/C	5.97	1.14	1.10-1.18	9.4×10^{-12}	rs2735940*	1.12	1.08-1.16	4.1×10^{-9}
5p13.1	*LINC00603*; *PTGER4*	rs7708610	5	40,102,443	A/G	35.64	1.04	1.02-1.06	1.5×10^{-5}	rs12514517*	1.06	1.04-1.08	3.8×10^{-9}
6p21.32	*HLA-B*; *MICA*; *MICB*; *NFKBIL1*; *TNF*	rs2516420	6	31,449,620	C/T	92.63	1.10	1.06-1.13	1.3×10^{-7}	rs9271695, rs116685461, rs116353863	1.12	1.08-1.16	2.0×10^{-10}
8q24.21	*MYC*	rs4313119	8	128,571,855	G/T	74.86	1.06	1.04-1.08	1.0×10^{-9}	rs6983267*, rs7013278	1.06	1.04-1.08	2.1×10^{-9}
12p13.32	*CCND2*	rs3217874	12	4,400,808	T/C	42.82	1.08	1.06-1.10	1.2×10^{-17}	rs3217810*, rs35808169*	1.06	1.04-1.08	2.4×10^{-9}
15q13.3	*GREM1*	rs17816465	15	33,156,386	A/G	20.55	1.07	1.04-1.09	6.8×10^{-9}	rs2293581*, rs12708491*	1.07	1.05-1.10	1.4×10^{-10}
20p12.3	*BMP2*	rs28488	20	6,762,221	T/C	63.88	1.06	1.04-1.08	2.6×10^{-11}	rs189583*, rs4813802*, rs994308	1.07	1.05-1.09	2.6×10^{-14}
20p12.3	*BMP2*	rs994308	20	6,603,622	C/T	59.39	1.08	1.06-1.10	4.8×10^{-18}	rs189583*, rs4813802*, rs28488	1.06	1.05-1.08	8.6×10^{-12}

Lead variant is the most associated variant at the locus in the conditional analysis. rsIDs based on NCBI dbSNP Build 150. Alleles are on the + strand. Chr.: Chromosome. RAF: Risk allele frequency, based on stage 2 data. OR, odds ratio estimates are for the risk allele. Conditioning variants are the lead variant of other conditionally independent association signals with $P < 1\times10^{-5}$ within 1-Mb of the new association signal. Because of extensive LD we used a 2-Mb distance for the MHC region (6p21.32). All lead variants for the new association signals are in linkage equilibrium with any previously reported CRC risk variants at the locus (r^2 <0.10).

*Indicates that the conditioning variant is either the index variant, or a variant in LD with the index variant reported in previous GWAS. Details and full results are provided in Supplementary Table 5.

ONLINE METHODS

Study samples.

After quality control (QC), this study included whole-genome sequencing (WGS) data for 1,439 colorectal cancer (CRC) cases and 720 controls from 5 studies, and GWAS array data for 58,131 CRC or advanced adenoma cases (3,674; 6.3% of cases) and 67,347 controls from 45 studies from GECCO, CORECT, and CCFR. The Stage 1 meta-analysis comprised existing genotyping data from 30 studies that were included in previously published CRC GWAS[13,18,22]. After QC, the Stage 1 meta-analysis included 34,869 cases and 29,051 controls. Study participants were predominantly of European ancestry (31,843 cases and 26,783 controls; 91.7% of participants). Because it was shown previously that the vast majority of known CRC risk variants are shared between Europeans and East Asians[17], we included 3,026 cases and 2,268 controls of East Asian ancestry to increase power for discovery. The Stage 2 meta-analysis comprised newly generated genotype data involving 4 genotyping projects and 22 studies. After QC, the Stage 2 meta-analysis included 23,262 cases and 38,296 controls, all of European ancestry. Studies, sample selection, and matching are described in the **Supplementary Text**. **Supplementary Table 1** provides details on sample numbers, and demographic characteristics of study participants. All participants provided written informed consent, and each study was approved by the relevant research ethics committee or institutional review board. Four normal colon mucosa biopsies for ATAC-seq were obtained from patients with a normal colon at colonoscopy at the Institut d'Investigació Biomèdica de Bellvitge (IDIBELL), Spain. Patients signed informed consent, and the protocol was approved by the Bellvitge Hospital Ethics Committee (Colscreen protocol PR084/16).

Whole-genome sequencing.

We performed low-pass WGS of 2,192 samples from 5 studies at the University of Washington Northwest Genomics Center (Seattle, WA, USA). Cases and controls were processed and sequenced together. Libraries were prepared with ThruPLEX DNA-seq kits (Rubicon Genomics) and paired-end sequencing performed using Illumina HiSeq 2500 sequencers. Reads were mapped to human reference genome (GRCh37 assembly) using Burrows-Wheeler aligner BWA v0.6.2[82]. Fold genomic coverage averaged 5.3× (range: 3.8-8.6×). We used the GotCloud population-based multi-sample variant calling pipeline[83] for post-processing of BAM files with

initial alignments, and to detect and call single nucleotide variants (SNVs) and short insertions and deletions (indels). After removing duplicated reads and recalibrating base quality scores, QC checks included sample contamination detection. Variants were jointly called across all samples. To identify high-quality sites, the GotCloud pipeline performs a two-step filtering process. First, lower quality variants are identified by applying individual variant quality statistic filters. Next, variants failing multiple filters are used as negative examples to train a support vector machine (SVM) classifier. Finally, we performed a haplotype-aware genotype refinement step via Beagle[84] and ThunderVCF[85] on the SVM-filtered VCF files. After further sample QC, we excluded samples with estimated DNA contamination >3% (16), duplicated samples (5) or related individuals (1), sex discrepancies (0), and samples with low concordance with GWAS array data (11). We checked for ancestry outliers by performing principal components analysis (PCA) after merging in data for shared, linkage disequilibrium (LD)-pruned SNVs for 1,092 individuals from the 1000 Genomes Project[86]. After QC, sequences were available for 1,439 CRC cases and 720 controls of European ancestry.

GWAS genotype data and quality control.

Details of genotyping and QC for studies included in the Stage 1 meta-analysis are described elsewhere[13,18,22]. **Supplementary Table 1** provides details of genotyping platforms used. Before association analysis, we pooled individual-level genotype data of all Stage 1 studies for a subset of SNPs to enable identification of unexpected duplicates and close relatives. We calculated identity by descent (IBD) for each pair of samples using KING-robust[87] and excluded duplicates and individuals that are second-degree or more closely related. As part of Stage 2, 28,805 individuals from 19 studies were newly genotyped on a custom Illumina array based on the Infinium OncoArray-500K[26] and a panel of 15,802 successfully manufactured custom variants (described in **Supplementary Text**). An additional 8,725 individuals from 5 studies were genotyped on the Illumina HumanOmniExpressExome-8v1-2 array. Genotyping and calling for both projects were performed at the Center for Inherited Disease Research (CIDR) at Johns Hopkins University. Genotypic data that passed initial QC at CIDR subsequently underwent QC at the University of Washington Genetic Analysis Center (UW GAC) using standardized methods detailed in Laurie *et al.*[88]. The median call rate for the custom Infinium OncoArray-500K data was 99.97%, and error rate estimated from 301 sample duplicate pairs was 9.99e-7. A

relatively low number of samples (246) had a missing call rate >2%, with the highest being 3.48%, and were included in analysis. For the HumanOmniExpressExome-8v1-2 data, median call rate was 99.96%, and the error rate estimated from 179 sample duplicate pairs was 2.65e-6. Thirty samples had a missing call rate >2%, with the highest being 3.79%, and were included in analysis. We excluded samples with discrepancies between reported and genotypic sex based on X chromosome heterozygosity and the means of sex chromosome probe intensities, unintentional duplicates, and close relatives defined as individuals that are second-degree or more closely related. After further excluding individuals of non-European ancestry as determined by PCA (see below), the custom OncoArray data included in analysis comprised 11,852 CRC cases and 11,895 controls, and the HumanOmniExpressExome-8v1-2 array data included in analysis comprised 4,439 CRC cases and 4,115 controls. Only variants passing QC were used for imputation. We excluded variants failing CIDR technical filters or UW GAC quality filters, which included missing call rate >2%, discordant calls in sample duplicates, and departures from Hardy-Weinberg equilibrium (HWE) (P <1e-4) based on European-ancestry controls. The Stage 2 analysis also included genotype data from the CORSA study (**Supplementary Text**). In total, 2,354 individuals were genotyped using the Affymetrix Axiom Genome-Wide Human CEU 1 Array. We called genotypes using the AxiomGT1 algorithm. All samples had missing call rate <3%. We excluded samples with discrepancies between reported and genotypic sex (20), close relatives defined as individuals that are second-degree or more closely related (94), as inferred using KING-robust[87], and individuals of non-European ancestry (6) as inferred from PCA. After QC, data included in analysis comprised 1,460 cases and 774 controls. Prior to phasing and imputation, we filtered out SNPs with missing call rate >2%, or HWE P <1e-4. Imputed genotype data were obtained from UK Biobank and QC and imputation are described elsewhere[89]. A nested case-control dataset was constructed as described in the **Supplementary Text**. We excluded individuals of non-European ancestry as inferred from PCA, and randomly dropped one individual from each pair that were more closely related than third-degree relatives as inferred using KING-robust. This resulted in excluding 137 samples. In total, 5,356 CRC (5,004) or advanced adenoma (352) cases and 21,407 matched controls were included in the replication analysis.

Principal components analysis.

After excluding close relatives, we performed PCA using PLINK1.9[90] on LD-pruned sets of autosomal SNPs obtained by removing regions with extensive long-range LD[91,92], SNPs with minor allele frequency (MAF) <5%, or HWE *P* <1e-4, or any missingness, and carrying out LD pruning using the PLINK option '-indep-pairwise 50 5 0.2'. To identify population outliers we merged in 1,092 individuals from 1000 Genomes Project Phase III and performed PCA using the intersection of variants[93].

Genotype imputation.

The 2,159 whole-genome sequences described above were used to create a phased imputation reference panel. After estimating haplotypes for all GWAS array data sets using SHAPEIT2[94], we used minimac3[95] to impute from this reference panel (19.6 million variants with minor allele count (MAC) >1) into the GWAS datasets described above. We also imputed to the Haplotype Reference Consortium (HRC) panel[25] (39.2 million variants) using the University of Michigan Imputation Server[95]. To improve imputation accuracy for Stage 1 data sets, phasing and imputation were performed after pooling studies/genotype projects that used the same, or very similar, genotyping platforms (**Supplementary Table 1**). For Stage 2, we performed phasing and imputation separately for each genotyping project data set and imputed to the HCR panel.

Statistical analyses.

Association testing of sequence data.

We tested variants with MAC ≥5 for CRC association using Firth's bias-reduced logistic regression as implemented in EPACTS (genome.sph.umich.edu/wiki/EPACTS) and adjusted for sex, age, study, and 3 principal components (PCs) calculated from an LD-pruned set of genotypes. We performed rare variant aggregate tests at the gene and enhancer level using the Mixed effects Score Test (MiST)[96]. This unified test is a linear combination between unidirectional burden and bidirectional variance component tests that performs best in terms of statistical power across a range of architectures[97].

Association and meta-analysis.

Stage 1 comprised two large mega-analyses of pooled individual-level genotype data sets (**Supplementary Table 12**). The four Stage 2 genotyping project data sets were analyzed

separately. Within each data set, variants with an imputation accuracy $r^2 \geq 0.3$ and MAC ≥ 50 were tested for CRC association using the imputed genotype dosage in a logistic regression model adjusted for age, sex, and study/genotyping project-specific covariates, including PCs to adjust for population structure (**Supplementary Table 12**). To account for residual confounding within CORSA, we tested association with each variant using a linear mixed model and kinship matrix calculated from the data, as implemented in EMMAX[98]. To enable meta-analysis, we then calculated approximate allelic log odds ratios (OR) and corresponding standard errors as described in Cook *et al.*[99].

Next, we combined association summary statistics across analyses via fixed-effects inverse variance-weighted meta-analysis. Because Wald tests can be notably anti-conservative for rare variant associations, we also performed likelihood ratio-based tests, followed by sample-size weighted meta-analysis, as implemented in METAL[100]. In total, 16,900,397 variants were analyzed. To examine residual population stratification, we inspected quantile-quantile plots of test statistics (**Supplementary Figure 8**), and calculated genomic control inflation statistics (λ_{GC}). λ_{GC} for the combined meta-analysis was 1.105, and for Stage 1 and 2 meta-analyses was 1.071 and 1.075, respectively. Because λ_{GC} increases with sample size for polygenic phenotypes, even in the absence of confounding biases[101], we investigated the effect of confounding due to residual population stratification using LD score regression[102]. Because of limitations of LD score regression, this analysis is restricted to common variants (MAF≥1%) for which λ_{GC} was 1.188 in the combined meta-analysis. The LD score regression intercept was 1.067, which is substantially less than λ_{GC}, indicating at most a small contribution of bias and that inflation in χ^2 statistics results mostly from polygenicity. We also calculated $\lambda_{1,000}$ which is the equivalent inflation statistic for a study with 1,000 cases and 1,000 controls[103]. For the combined meta-analysis, λ_{1000} was 1.004 and for both Stage 1 and 2 meta-analyses this was 1.003.

Significance threshold for the replication genotyping experiment.

To protect against probe design failure, we built redundancy into the custom genotyping panel by including LD proxies of independently associated variants selected for follow-up. To determine the number of independent tests, we performed LD clumping of the 9,198 analyzed variants that were selected for replication genotyping based on the Stage 1 meta-analysis, and that survived

filters described above. Using an r^2 threshold of 0.1 this translated to 6,438 independent tests and a Bonferroni significance threshold of $0.05/6{,}438=7.8\times10^{-6}$.

Conditional and joint multiple-variant analysis.

To identify additional distinct association signals at CRC loci, we performed a series of conditional meta-analyses. At each locus attaining $P<5\times10^{-8}$, we included the genotype dosage for the variant showing the strongest statistical evidence for association in the region in the combined meta-analysis, as an additional covariate in the respective logistic regression models. Association summary statistics for each variant in the region were then combined across studies by a fixed-effects meta-analysis. If at least one association signal attained a significance level of $P<1\times10^{-5}$ in this meta-analysis, we performed a second round of conditional meta-analysis, adding the variant showing the strongest statistical evidence for association in the region in the first round of conditional meta-analysis as a covariate to the logistic regression models used in the first round. We repeated this procedure and kept adding variants to the model until no additional variants at the locus attained $P<1\times10^{-5}$. Finally, we performed a joint multiple-variant analysis in which we jointly estimated the effects of variants selected in each step and tested for each variant whether the *P*-value from the joint multiple-variant analysis (P_J) was $<1\times10^{-5}$. Analyses were performed on 2-Mb windows centered on the most associated variant in the unconditional analysis. If windows overlapped, we performed the analysis on the collapsed genomic region. Because of extensive LD, we used a 4-Mb window for the MHC region.

Definition of known loci.

We compiled a list of 62 previously reported genome-wide significant CRC association signals from the literature (**Supplementary Table 3**). Because of improved power and coverage of our study, we identified the most associated variant at each signal, and used these lead variants for further analyses, rather than the previously reported index variant.

Refinement of association signals.

To refine new association signals, we constructed credible sets that were 99% likely, based on posterior probability, to contain the causal disease-associated SNP[104]. In brief, for each distinct signal, we retained a candidate set of variants by identifying all analyzed variants with $r^2 \geq 0.1$

with the most associated variant within a 2-Mb window centered on the most associated variant. We calculated approximate Bayes' factors (ABF)[105] for each variant as:

$$ABF = \sqrt{1-r}\, e^{rz^2/2}$$

where $r = 0.04/(\text{s.e.}^2+0.04)$, $z = \beta/\text{s.e.}$, and β and s.e. are the log OR estimate and its standard error from the combined meta-analysis. For loci with multiple distinct signals, results are based on conditional meta-analysis, adjusting for all other index variants in the region. We then calculated the posterior probability of being causal as ABF/T where T is the sum of ABF values over all candidate variants. Next, variants were ranked in decreasing order by posterior probabilities and the 99% credible set was obtained by including variants with the highest posterior probabilities until the cumulative posterior probability ≥99%.

Functional genomic annotation.

To nominate variants for future laboratory follow-up, we performed bioinformatic analysis at each new signal using our functional annotation database, and a custom UCSC analysis data hub. Using ANNOVAR[106]**,** we annotated lead variants and variants in LD (r^2 ≥0.4) with the lead variant, relative to features pertaining to i) gene-centric function (PolyPhen2[107]), ii) genome-wide functional prediction scores (CADD[108], DANN[109], EigenPC[110]), iii) disease relatedness (GWAS catalog[46]), and iv) CRC-relevant regulatory functions (enhancer, repressor, DNA accessible, and transcription factor binding site (TFBS)[111,112]; **Supplementary Table 13**). **Supplementary Table 8** summarizes variant annotations relative to the CCDS Project[113], and reference genome GRCh37. Variants were maintained in **Supplementary Table 8** if they met any of the following conditions: DANN score ≥0.9, CADD phred score ≥20, Eigen-PC phred score ≥17, PolyPhen2 "probably damaging", "stop loss", "stop gain", "splicing", or were positioned in a predicted regulatory element. We visually inspected loci overlapping with CRC-relevant functional genomic annotations. Variants positioned in enhancers with aberrant CRC activity were identified by comparing epigenomes of non-diseased colorectal tissues/colonic crypt cells to epigenomes of primary CRC cell lines (data accessible at NCBI GEO database, accession GSE77737). We prioritized target genes for loci with predicted regulatory function. Evidence suggests that Topological Association Domains (TADs) can be used to map physical

boundaries on gene promoter interactions with distal regulatory elements[114–116]. As such, we used GMI12878 Hi-C Chromosome Conformation Capture data to identify gene promoters that were in the same TADs as risk loci using the WashU Epigenome Browser (https://epigenomegateway.wustl.edu/). Genes in this list were further prioritized based on biological relevancy and expression quantitative trait loci (eQTL) data from Genotype-Tissue Expression (GTEx)[117] using HaploReg v4.1[118].

ATAC-seq assay.

We generated high resolution maps of DNA accessible regions in normal colon mucosa samples using the Assay for Transposase-Accessible Chromatin using sequencing (ATAC-seq). Using the updated omni-ATAC protocol for archival samples, we performed ATAC-seq in four colon mucosa biopsies from the ICO-biobank taken from participants undergoing screening at IDIBELL, Spain. Biopsies were cryopreserved by slow freezing using a solution of 10% DMSO, 90% media, and Mr. Frosty Cryo 1°C Freezing Containers (Thermo Scientific). ATAC-seq was implemented as prescribed with two exceptions. Instead of dounce homogenizer we used a tissue lyser and stainless bead system, pulverizing at 40Hz for 2 mins and pulsing at 50Hz for 10-20 seconds. Secondly, Illumina library quantification was performed using picogreen quantitation and TapeStation instead of KAPA quantitative qPCR. Libraries were sequenced to an average of 25M paired end reads using Illumina HiSeq 2500. The ENCODE data processing pipeline was implemented (https://github.com/kundajelab/atac_dnase_pipelines) aligning to hg19[119]. QC results are summarized in **Supplementary Table 14**.

Regulatory and functional information enrichment analysis.

We used GARFIELD[74] to identify cell types, tissues, and functional genomic features relevant to CRC risk. This method tests for enrichment of association in features primarily extracted from ENCODE and Roadmap Epigenomics Project data, while accounting for sources of confounding, including LD. We applied default settings and used the author-supplied data which is suitable for analysis of GWAS results based on European-ancestry individuals.

Pathway and gene set enrichment analysis.

We used MAGENTA to test predefined gene sets (e.g., KEGG pathways) for enrichment for CRC risk associations[75]. We used combined meta-analysis results as input and applied default settings which included removing genes that fall in the MHC region from analysis. Enrichment was tested at two gene *P*-value cutoffs: 95th and 75th percentiles of all gene *P*-values in the genome.

Estimation of contribution of rare variants to heritability.

We used the LD- and MAF-stratified component GREML (GREML-LDMS) method as implemented in GCTA[76] to estimate the proportion of variation in liability for CRC explained by all imputed autosomal variants (i.e., estimate of narrow-sense heritability h_g^2), and the proportion contributed by rare variants (MAF ≤1%). Because of computational limitations we analyzed a subset of 11,895 cases and 14,659 controls imputed to our WGS panel. We analyzed individual-level data for 17,649,167 imputed variants with MAC >3 and HWE test $P \geq 10^{-6}$. Following Yang *et al.*[76], we did not filter on imputation quality. In brief, we stratified variants into groups based on MAF (boundaries at 0.001, 0.01, 0.1, 0.2, 0.3, 0.4) and mean LD score (boundaries at quartiles) calculated as described in Yang *et al.*[76]. We then calculated genetic relationship matrices (GRMs) for each of these 28 variant partitions and jointly estimated variance components for these partitions, adjusting for age, sex, study, genotyping batch, and three genotype PCs. From the variance component estimates and their variance-covariance matrix we estimated the contribution of rare variants (MAF ≤1%) and common variants (MAF >1%), and calculated standard errors using the delta method. We tested significance of the contribution of rare variants using a likelihood ratio test. To calculate heritability on the underlying liability scale we interpreted *K* as lifetime risk[120] and used an estimate of 4.3% (Surveillance, Epidemiology, and End Results Program (SEER) Cancer Statistics, 2011-2013).

Familial relative risk explained by genetic variants.

We assumed a multiplicative model within and between variants and calculated the proportion of familial relative risk (RR) explained by a given set of genetic variants as $\frac{\sum_i log\ \lambda_i}{log\ \lambda_0}$, where λ_0 is the overall familial RR to first-degree relatives of cases. λ_i is the familial RR due to variant i calculated as $\lambda_i = \frac{p_i r_i^2 + q_i}{(p_i r_i + q_i)^2}$, where p_i is the risk allele frequency for variant i, $q_i = 1 - p_i$, and r_i

is the estimated per allele OR[9,121]. We adjusted the OR estimates of new association signals for winner's curse following Zhong and Prentice[78]. We represented previously identified association signals by the variant showing the strongest statistical evidence of association in the combined meta-analysis, and assumed that winner's curse was negligible. We assumed λ_0 to be 2.2[122]. Using the delta method, we computed the variance for the proportion of familial RR as follows:

$$\sum_i Var(r_i)\,[\frac{1}{log\ \lambda_0}\frac{1}{\lambda_i}\frac{2p_iq_i(r_i-1)}{(p_ir_i+q_i)^3}]^2.$$

Absolute risk of CRC incidence and starting age of first screening.

We constructed a polygenic risk score (PRS) as a weighted sum of expected risk allele frequency for common genetic variants, using the per allele OR for each variant as weights. OR estimates for newly discovered variants were adjusted for winner's curse to avoid potential inflation[78]. Assuming all genetic variants are independent, let X denote a PRS constructed based on K variants: $X = \sum_{i=1}^{K} \widehat{\beta_i} Z_i$, where $\widehat{\beta_i}\ and\ Z_i$ are the estimated OR and the number of risk alleles for variant i. We assumed X follows a normal distribution $N(\mu, \sigma^2)$, where the estimates of mean and variance are computed as following:

$$\hat{\mu} = \sum_{i=1}^{K} \widehat{\beta_i}\times 2\times \widehat{p_i} \text{ and } \widehat{\sigma^2} = \sum_{i=1}^{K} \widehat{\beta_i^2}\times 2\times \widehat{p_i}\times(1-\widehat{p_i}),$$

where $\widehat{p_i}$ is the risk allele frequency for variant $i = 1, \cdots, K$. Then the baseline hazard at each age t, $\widehat{\lambda_0}(t)$, is computed as following:

$$\widehat{\lambda_0}(1) = \lambda^*(1)\frac{\int f(x)\,dx}{\int e^x f(x)dx}$$

$$\widehat{\lambda_0}(t) = \lambda^*(t)\frac{\int exp\,(-\sum_{i=1}^{t-1}\widehat{\lambda_0}(i)\,e^x)\,f(x)dx}{\int exp\,(-\sum_{i=1}^{t-1}\widehat{\lambda_0}(i)\,e^x)\,e^x f(x)dx} \text{ for } t = 2, \cdots, 100,$$

and $\lambda^*(t)$ are the incidence rates for non-Hispanic whites who have not taken an endoscopy before, derived from population incidence rates during 1992-2005 from the SEER Registry. Using these baseline hazard rates, we estimated the 10-year absolute risk of developing CRC given age and a PRS as previously described[123]. By setting a risk threshold as the average of the 10-year CRC risk for a 50-year old man (1.25%) and woman (0.68%), i.e., (1.25%+0.68%)/2=0.97%, who have not previously received an endoscopy[124], we estimated the

recommended starting age of first screening given the PRS. Variants and OR estimates used in these analyses are given in **Supplementary Table 15**.

Data availability.

All whole-genome sequence data have been deposited at the database of Genotypes and Phenotypes (dbGaP), which is hosted by the U.S. National Center for Biotechnology Information (NCBI), under accession number phs001554.v1.p1. All custom Infinium OncoArray-500K array data for the studies in the Stage 2 meta-analysis have been deposited at dbGaP under accession number phs001415.v1.p1. All Illumina HumanOmniExpressExome-8v1-2 array data for the studies in the Stage 2 meta-analysis have been deposited at dbGaP under accession number phs001315.v1.p1. Genotype data for the studies included in the Stage 1 meta-analysis have been deposited at dbGaP under accession number phs001078.v1.p1. The UK Biobank resource was accessed through application number 8614.

Reporting Summary.

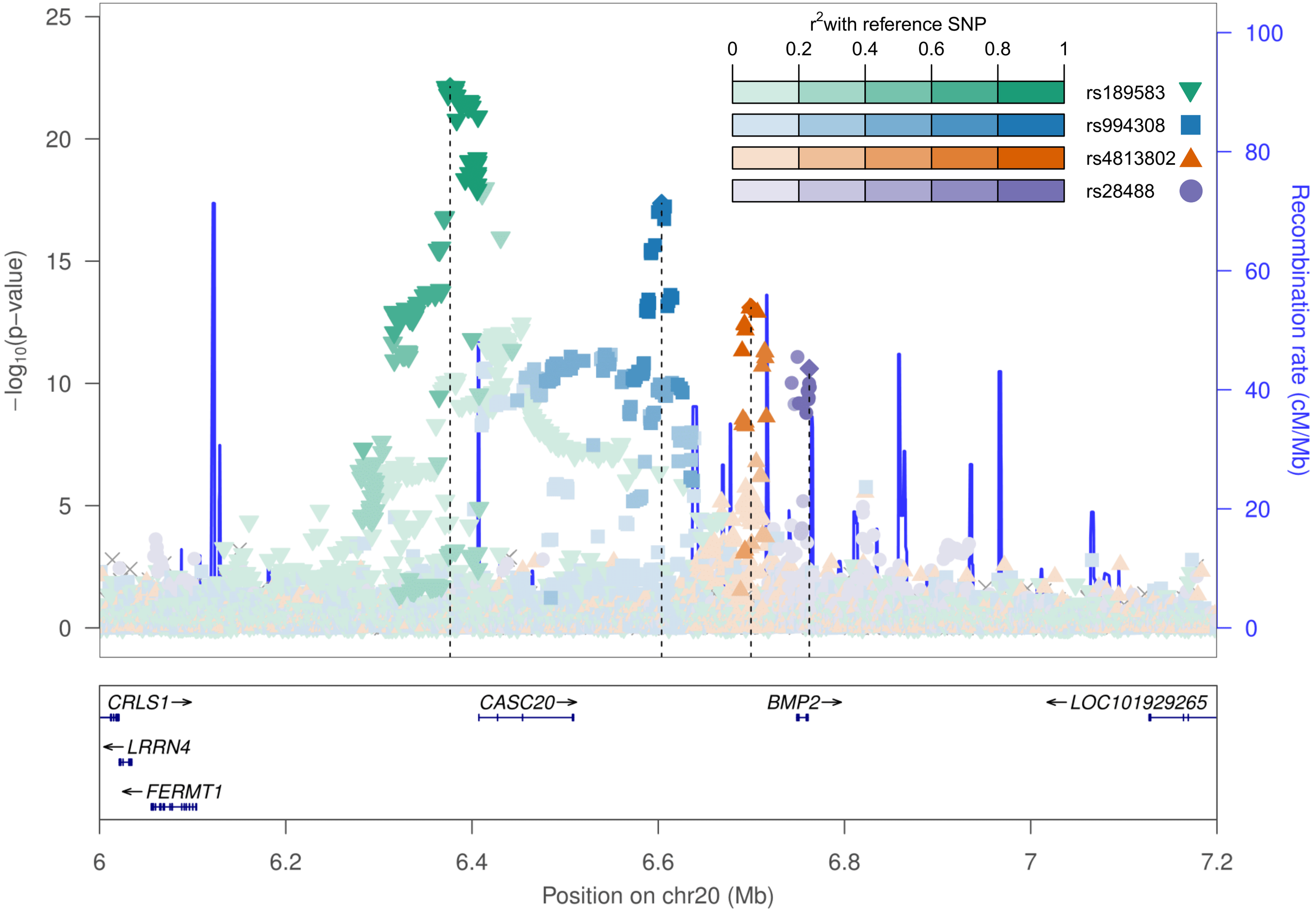
Further information on experimental design is available in the Life Sciences Reporting Summary linked to this article.

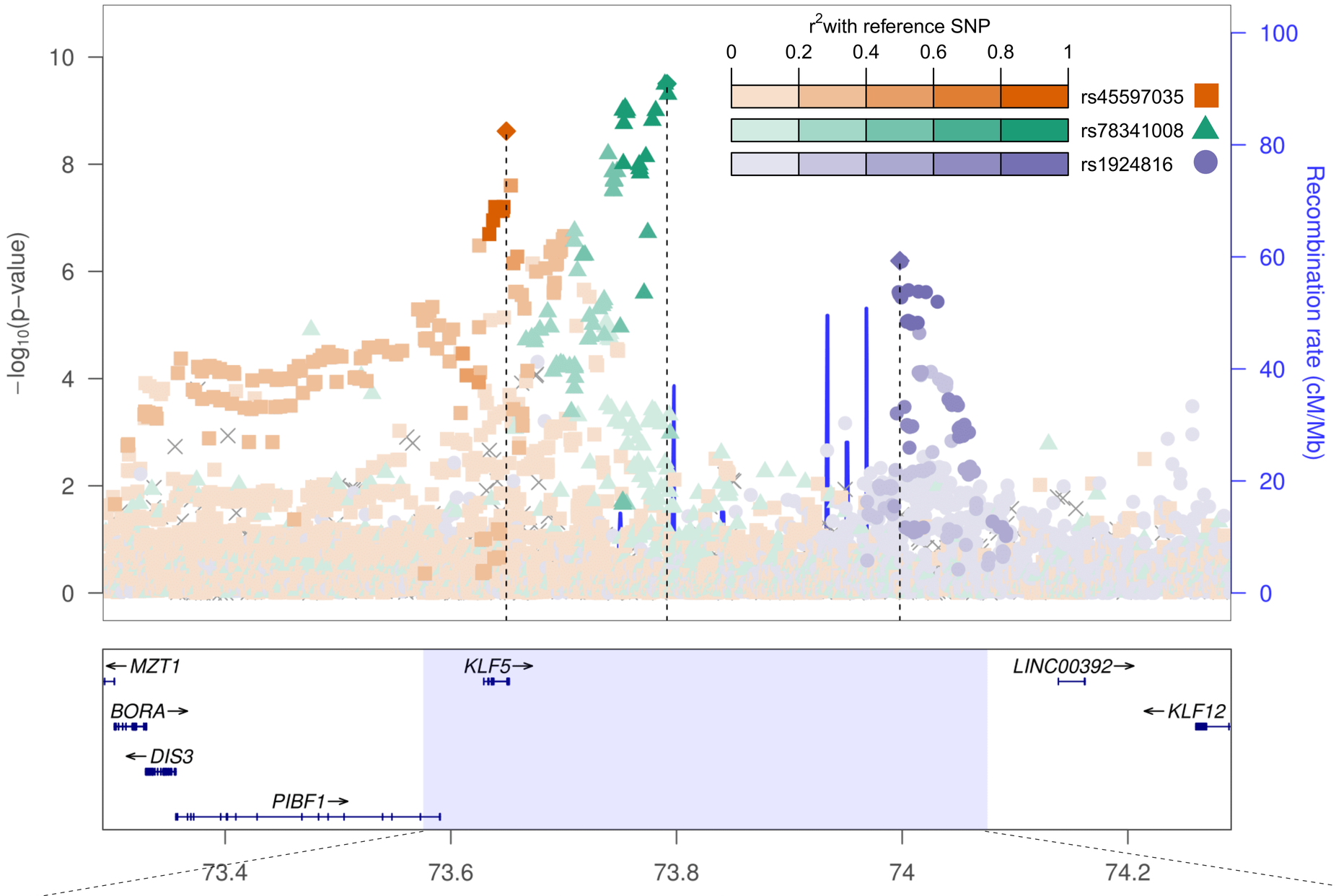
METHODS-ONLY REFERENCES

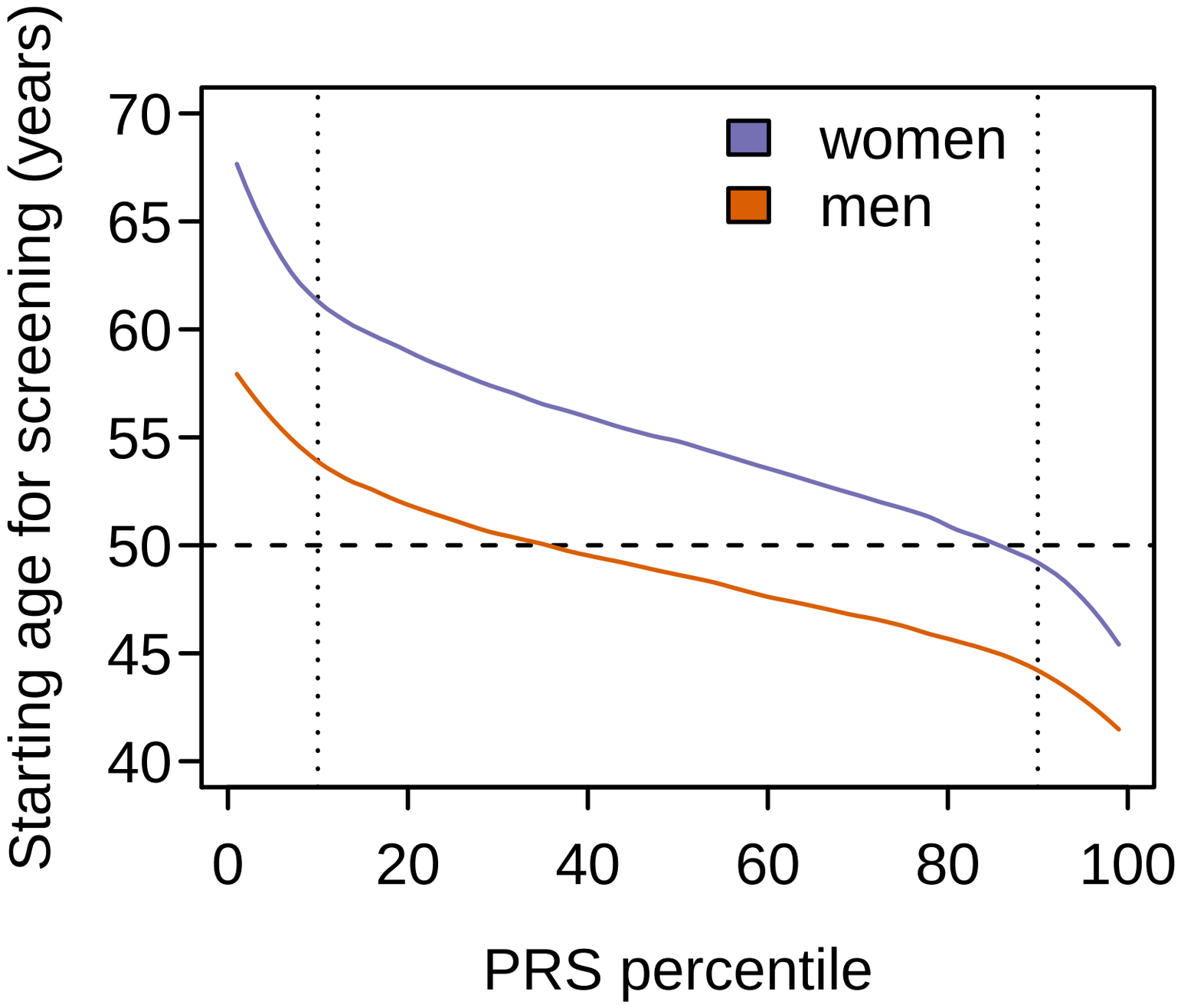
82. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25,** 1754–1760 (2009).
83. Jun, G., Wing, M. K., Abecasis, G. R. & Kang, H. M. An efficient and scalable analysis framework for variant extraction and refinement from population-scale DNA sequence data. *Genome Res* **25,** 918–925 (2015).
84. Browning, B. L. & Yu, Z. Simultaneous genotype calling and haplotype phasing improves genotype accuracy and reduces false-positive associations for genome-wide association studies. *Am J Hum Genet* **85,** 847–861 (2009).
85. Li, H. A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. *Bioinformatics* **27,** 2987–2993 (2011).
86. 1000 Genomes Project Consortium *et al.* A map of human genome variation from population-scale sequencing. *Nature* **467,** 1061–1073 (2010).
87. Manichaikul, A. *et al.* Robust relationship inference in genome-wide association studies. *Bioinformatics* **26,** 2867–2873 (2010).
88. Laurie, C. C. *et al.* Quality control and quality assurance in genotypic data for genome-

wide association studies. *Genet Epidemiol* **34,** 591–602 (2010).
89. Bycroft, C. *et al.* Genome-wide genetic data on ~500,000 UK Biobank participants. *BioRxiv* (2017). doi:10.1101/166298
90. Chang, C. C. *et al.* Second-generation PLINK: rising to the challenge of larger and richer datasets. *Gigascience* **4,** 7 (2015).
91. Price, A. L. *et al.* Long-range LD can confound genome scans in admixed populations. *Am J Hum Genet* **83,** 132–135 (2008).
92. Weale, M. E. Quality control for genome-wide association studies. *Methods Mol Biol* **628,** 341–372 (2010).
93. 1000 Genomes Project Consortium *et al.* An integrated map of genetic variation from 1,092 human genomes. *Nature* **491,** 56–65 (2012).
94. Delaneau, O., Howie, B., Cox, A. J., Zagury, J.-F. & Marchini, J. Haplotype estimation using sequencing reads. *Am J Hum Genet* **93,** 687–696 (2013).
95. Das, S. *et al.* Next-generation genotype imputation service and methods. *Nat Genet* **48,** 1284–1287 (2016).
96. Sun, J., Zheng, Y. & Hsu, L. A unified mixed-effects model for rare-variant association in sequencing studies. *Genet Epidemiol* **37,** 334–344 (2013).
97. Moutsianas, L. *et al.* The power of gene-based rare variant methods to detect disease-associated variation and test hypotheses about complex disease. *PLoS Genet* **11,** e1005165 (2015).
98. Kang, H. M. *et al.* Variance component model to account for sample structure in genome-wide association studies. *Nat Genet* **42,** 348–354 (2010).
99. Cook, J. P., Mahajan, A. & Morris, A. P. Guidance for the utility of linear models in meta-analysis of genetic association studies of binary phenotypes. *Eur J Hum Genet* **25,** 240–245 (2017).
100. Willer, C. J., Li, Y. & Abecasis, G. R. METAL: fast and efficient meta-analysis of genomewide association scans. *Bioinformatics* **26,** 2190–2191 (2010).
101. Yang, J. *et al.* Genomic inflation factors under polygenic inheritance. *Eur J Hum Genet* **19,** 807–812 (2011).
102. Bulik-Sullivan, B. K. *et al.* LD Score regression distinguishes confounding from polygenicity in genome-wide association studies. *Nat Genet* **47,** 291–295 (2015).
103. Michailidou, K. *et al.* Large-scale genotyping identifies 41 new loci associated with breast cancer risk. *Nat Genet* **45,** 353–61, 361e1 (2013).
104. Wellcome Trust Case Control Consortium *et al.* Bayesian refinement of association signals for 14 loci in 3 common diseases. *Nat Genet* **44,** 1294–1301 (2012).
105. Wakefield, J. A Bayesian measure of the probability of false discovery in genetic epidemiology studies. *Am J Hum Genet* **81,** 208–227 (2007).
106. Wang, K., Li, M. & Hakonarson, H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Res* **38,** e164 (2010).
107. Adzhubei, I., Jordan, D. M. & Sunyaev, S. R. Predicting functional effect of human missense mutations using PolyPhen-2. *Curr Protoc Hum Genet* **Chapter 7,** Unit7.20 (2013).
108. Kircher, M. *et al.* A general framework for estimating the relative pathogenicity of human genetic variants. *Nat Genet* **46,** 310–315 (2014).
109. Quang, D., Chen, Y. & Xie, X. DANN: a deep learning approach for annotating the pathogenicity of genetic variants. *Bioinformatics* **31,** 761–763 (2015).

110. Ionita-Laza, I., McCallum, K., Xu, B. & Buxbaum, J. D. A spectral approach integrating functional genomic annotations for coding and noncoding variants. *Nat Genet* **48,** 214–220 (2016).
111. Roadmap Epigenomics Consortium *et al.* Integrative analysis of 111 reference human epigenomes. *Nature* **518,** 317–330 (2015).
112. Corradin, O. *et al.* Combinatorial effects of multiple enhancer variants in linkage disequilibrium dictate levels of gene expression to confer susceptibility to common traits. *Genome Res* **24,** 1–13 (2014).
113. Pruitt, K. D. *et al.* The consensus coding sequence (CCDS) project: Identifying a common protein-coding gene set for the human and mouse genomes. *Genome Res* **19,** 1316–1323 (2009).
114. Harmston, N. *et al.* Topologically associating domains are ancient features that coincide with Metazoan clusters of extreme noncoding conservation. *Nat Commun* **8,** 441 (2017).
115. Berlivet, S. *et al.* Clustering of tissue-specific sub-TADs accompanies the regulation of HoxA genes in developing limbs. *PLoS Genet* **9,** e1004018 (2013).
116. Hu, Z. & Tee, W.-W. Enhancers and chromatin structures: regulatory hubs in gene expression and diseases. *Biosci Rep* **37,** (2017).
117. GTEx Consortium. The Genotype-Tissue Expression (GTEx) project. *Nat Genet* **45,** 580–585 (2013).
118. Ward, L. D. & Kellis, M. HaploReg: a resource for exploring chromatin states, conservation, and regulatory motif alterations within sets of genetically linked variants. *Nucleic Acids Res* **40,** D930-4 (2012).
119. Landt, S. G. *et al.* ChIP-seq guidelines and practices of the ENCODE and modENCODE consortia. *Genome Res* **22,** 1813–1831 (2012).
120. Witte, J. S., Visscher, P. M. & Wray, N. R. The contribution of genetic variants to disease depends on the ruler. *Nat Rev Genet* **15,** 765–776 (2014).
121. Cox, A. *et al.* A common coding variant in CASP8 is associated with breast cancer risk. *Nat Genet* **39,** 352–358 (2007).
122. Johns, L. E. & Houlston, R. S. A systematic review and meta-analysis of familial colorectal cancer risk. *Am J Gastroenterol* **96,** 2992–3003 (2001).
123. Hsu, L. *et al.* A model to determine colorectal cancer risk using common genetic susceptibility loci. *Gastroenterology* **148,** 1330–9.e14 (2015).
124. Jeon, J. et al. Determining risk of colorectal cancer and starting age of screening based on lifestyle, environmental, and genetic factors. Gastroenterology 154, 2152–2164.e19 (2018).


r2with reference SNP
0
0.2
0.4
0.6
0.8
1
rs189583
rs994308
rs4813802
rs28488
−log10(p−value)
25
20
15
10
5
0
100
80
60
40
20
0
Recombination rate (cM/Mb)
CRLS1→
CASC20→
BMP2→
←LOC101929265
←LRRN4
←FERMT1
6
6.2
6.4
6.6
6.8
7
7.2
Position on chr20 (Mb)


r2with reference SNP
0
0.2
0.4
0.6
0.8
1
rs45597035
rs78341008
rs1924816
−log10(p−value)
Recombination rate (cM/Mb)
MZT1
KLF5
LINC00392
BORA
KLF12
DIS3
PIBF1
73.4
73.6
73.8
74
74.2
Position on chr13 (Mb)
Window Position
chr13:73,575,577-74,075,561 (499,985 bp)
Scale
100 kb
chr13:
73,650,000
73,700,000
73,750,000
73,800,000
73,850,000
73,900,000
73,950,000
74,000,000
74,050,000
PIBF1
KLF5
NormalCrypt Enhancer (n=4)
NormalCrypt Dhs (n=4)
NormalMucosa Atac-seq (n=4)
CRC Enhancer (n=32)
CRC Dhs (n=3)
LD proxies
Credible Set
Super enhancer
CRC cell lines (n=32)
Enhancers from Zhang et al 2018
e1
e3
e4


Starting age for screening (years)
70
65
60
55
50
45
40
0
20
40
60
80
100
PRS percentile
women
men